DYNAMICAL SYSTEMS AND NON-HERMITIAN ITERATIVE EIGENSOLVERS

MARK EMBREE* AND RICHARD B. LEHOUCQ†

1 September 2007

Abstract. This paper describes an equivalence between certain discrete dynamical systems and a class of iterations for the non-Hermitian eigenvalue problem. The identification of these discrete iterations as approximate solutions of ordinary differential equations reveals important geometric properties that, for example, provide insight into the role of orthogonality and bi-orthogonality. We demonstrate that the continuous systems possess a quadratic invariant and describe the drift of this conserved quantity under discretization; this invariant plays an important role in the convergence and stability of the discrete iteration. We also investigate the effect of preconditioning on the convergence and stability of the continuous system and its discretization.

Key words. dynamical systems, inverse iteration, preconditioning, eigenvalues, invariants

AMS subject classifications. 15A18, 37C10, 65F15, 65L20

1. Introduction. Suppose we seek a small number of eigenvalues of the non-Hermitian matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, having at our disposal a nonsingular matrix $\mathbf{N} \in \mathbb{C}^{n\times n}$ that will serve as a preconditioner for $\mathbf{A}$. Over recent years there has been a growing interest in the computation of eigenvalues via *preconditioned inverse iteration*. Given a starting vector $\mathbf{p}_0 \in \mathbb{C}^n$, compute

$$\mathbf{p}_{j+1} = \mathbf{p}_j + \mathbf{N}^{-1}(\theta_j - \mathbf{A})\mathbf{p}_j, \tag{1.1}$$

where $\theta_j - \mathbf{A}$ is shorthand for $\mathbf{I}\theta_j - \mathbf{A}$, and

$$\theta_j = \frac{(\mathbf{A}\mathbf{p}_j, \mathbf{p}_j)}{(\mathbf{p}_j, \mathbf{p}_j)}$$

for some inner product $(\cdot, \cdot)$. See, for example, [14, 15] and the references therein for the analysis of such iterations for Hermitian positive definite $\mathbf{A}$. With the ideal preconditioner $\mathbf{N} = \mathbf{A}$, this method reduces to (scaled) inverse iteration:

$$\mathbf{p}_{j+1} = \mathbf{A}^{-1}\mathbf{p}_j\theta_j.$$

Hence if $\mathbf{N}$ approximates $\mathbf{A}$, at least with respect to some eigenspace, we might expect that the sequence $\{\mathbf{p}_j\}$ (with suitable normalization) converges to an eigenvector of $\mathbf{A}$. The method (1.1) is but one example of a broader class of methods, and one objective of the present study is to develop a framework for the classification and analysis of this family.

The iteration (1.1) can be viewed as the forward Euler discretization of the autonomous nonlinear differential equation

$$\dot{\mathbf{p}} = \mathbf{N}^{-1}\Big(\mathbf{p}\frac{(\mathbf{A}\mathbf{p}, \mathbf{p})}{(\mathbf{p}, \mathbf{p})} - \mathbf{A}\mathbf{p}\Big) \tag{1.2}$$

*Department of Computational and Applied Mathematics, Rice University, 6100 Main Street – MS 134, Houston, TX 77005–1892 (embree@rice.edu). Supported by U.S. Department of Energy grant DE-FG03-02ER25531 and National Science Foundation grant DMS-CAREER-0449973.

†Sandia National Laboratories, P.O. Box 5800, MS 1110, Albuquerque, NM 87185–1110 (rblehou@sandia.gov). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy under contract DE-AC04-94AL85000.

with a unit step size. The nonzero steady states of this system correspond to (right) eigenvectors of $\mathbf{A}$, and hence one might attempt to compute eigenvalues by driving this differential equation to steady state as swiftly as possible. In this paper we examine connections between simple eigenvalue iterations and related differential equations.

There exists a longstanding association of eigenvalue iterations with differential equations [1, 6, 9, 12], with notable examples including Rayleigh quotient gradient flow (see, e.g., [18]), connections between the QR algorithm for dense eigenproblems and Toda flow [19, 28], and more general "isospectral flows" [31]. Although not developed as an algorithm for the algebraic eigenvalue problem, the Car–Parrinello method [4] determines the Kohn–Sham eigenstates from a second order ordinary differential equation, Newton's equations of motion (see [24, p 1086] for a formulation using the unpreconditioned (1.2)). The heavy ball optimization method [25] also formulates the minimum of the Rayleigh quotient via a second order ordinary differential equation. The paper [3] computes the ground state solution of Bose–Einstein condensates by using a normalized gradient flow discretized by several time integration schemes. The Kohn–Sham eigenstates and Bose–Einstein condensates give rise to self-adjoint nonlinear eigenvalue problems.

The differential equation (1.2) enjoys a distinguished property. Suppose that $\mathbf{p}$ solves (1.2), $\mathbf{N}$ is self-adjoint and invertible, and $\theta = (\mathbf{p},\mathbf{p})^{-1}(\mathbf{Ap},\mathbf{p})$. Then for all t,

$$\begin{aligned}\frac{d}{dt}(\mathbf{p},\mathbf{Np}) &= \big(\mathbf{N}^{-1}(\mathbf{p}\theta-\mathbf{Ap}),\mathbf{Np}\big)+\big(\mathbf{p},\mathbf{N}\mathbf{N}^{-1}(\mathbf{p}\theta-\mathbf{Ap})\big)\\ &= (\mathbf{p}\theta,\mathbf{p})-(\mathbf{Ap},\mathbf{p})+(\mathbf{p},\mathbf{p}\theta)-(\mathbf{p},\mathbf{Ap})\\ &= 0. \end{aligned} \tag{1.3}$$

Thus $(\mathbf{p},\mathbf{Np})$ is an *invariant* (or *first integral*), as its value is independent of time; see [12, §1.3] for a discussion of the unpreconditioned case ($\mathbf{N}=\mathbf{I}$), and, e.g., [2, 11] for a general introduction to invariant theory and geometric integration. When $\mathbf{N}=\mathbf{I}$, we call (1.1) an orthogonal correction method.

The invariant describes a manifold in n-dimensional space, $(\mathbf{p},\mathbf{Np})=(\mathbf{p}_0,\mathbf{Np}_0)$, on which the solution to the differential equation with $\mathbf{p}(0)=\mathbf{p}_0$ must fall. Simple discretizations such as Euler's method do not typically respect such invariants, and thus solutions can drift from the manifold. Our goal is to explain the relationship between convergence and stability for the continuous and discrete dynamical systems. In particular, the quadratic invariant is a crucial property of the continuous system, and it plays an important role in the convergence theory of the corresponding forward Euler discretization.

These discretizations (equivalently, preconditioned iterations for large scale non-Hermitian eigenvalue problems) are especially appealing in situations where a shift-invert Arnoldi method is intractable due to the onerous cost of preconditioned iterative methods used in inexact inner iterations; see [17, 30] for examples.

2. Dynamical systems and invariant manifolds. We are interested in generalizations of the simple preconditioned iteration that are appropriate for non-Hermitian matrix pencils, and properties of the dynamical systems from which such iterations arise. In this section and the next we shall focus on unpreconditioned iterations, $\mathbf{N}=\mathbf{I}$, before considering the influence of preconditioners in Section 5.

Let $\mathbf{A},\mathbf{B}\in\mathbb{C}^{n\times n}$ be matrices (with constant entries). For the generalized eigenvalue problem $\mathbf{Ax}=\mathbf{Bx}\lambda$ with $\mathbf{N}=\mathbf{I}$, the preconditioned dynamical system (1.2)

expands to

$$\dot{\mathbf{p}} = \mathbf{B}\mathbf{p}\theta - \mathbf{A}\mathbf{p}$$

for appropriate $\theta = \theta(t)$. This system suggests a development from iterations for the single vector $\mathbf{p} \in \mathbb{C}^n$ to iterations for subspaces, $\operatorname{Ran}\mathbf{P}$, where $\mathbf{P} \in \mathbb{C}^{n\times k}$:

$$\dot{\mathbf{P}} = \mathbf{B}\mathbf{P}\mathbf{L} - \mathbf{A}\mathbf{P};$$

we shall address the choice of $\mathbf{L} \in \mathbb{C}^{k\times k}$ momentarily. (Quantities such as $\mathbf{L}$ are t-dependent unless explicitly stated otherwise; we typically suppress the t argument to simplify notation.)

For non-Hermitian problems one might simultaneously evolve an equation for the adjoint to obtain approximations to the left eigenspace, which suggests the system

$$\begin{aligned} \dot{\mathbf{P}} &= \mathbf{B}\mathbf{P}\mathbf{L} - \mathbf{A}\mathbf{P} \\ \dot{\mathbf{Q}} &= \mathbf{B}^*\mathbf{Q}\mathbf{M}^* - \mathbf{A}^*\mathbf{Q}, \end{aligned} \tag{2.1}$$

with initial conditions $\mathbf{P}(0) = \mathbf{P}_0$ and $\mathbf{Q}(0) = \mathbf{Q}_0$, where $\mathbf{P}, \mathbf{Q} \in \mathbb{C}^{n\times k}$, and $\mathbf{L}, \mathbf{M} \in \mathbb{C}^{k\times k}$. Here $\cdot^*$ denotes the conjugate transpose and $(\cdot,\cdot)$ the standard Euclidean inner product (though this analysis generalizes readily to arbitrary inner products). The choice we make for the time-dependent $\mathbf{L}, \mathbf{M} \in \mathbb{C}^{k\times k}$ can potentially couple $\mathbf{P}$ and $\mathbf{Q}$. At steady state,

$$\mathbf{B}\mathbf{P}\mathbf{L} = \mathbf{A}\mathbf{P}, \quad \mathbf{B}^*\mathbf{Q}\mathbf{M}^* = \mathbf{A}^*\mathbf{Q},$$

and hence, provided $\mathbf{P}$ and $\mathbf{Q}$ have full column rank, the eigenvalues of $\mathbf{L}$ and $\mathbf{M}$ are included in the spectrum of the pencil $\mathbf{A} - \lambda\mathbf{B}$, while the columns of $\mathbf{P}$ and $\mathbf{Q}$ span right- and left-invariant subspaces of the same pencil. We shall motivate the choice of the matrices $\mathbf{L}$ and $\mathbf{M}$ through generalizations of the invariant discussed in the introduction.

The following notation facilitates the analysis of these subspace iterations.

DEFINITION 2.1. *Given $\mathbf{P}, \mathbf{Q} \in \mathbb{C}^{n\times k}$, define $(\mathbf{P}, \mathbf{Q}) = \mathbf{Q}^*\mathbf{P} \in \mathbb{C}^{k\times k}$, i.e., the (i, j) entry of $(\mathbf{P}, \mathbf{Q})$ satisfies $(\mathbf{P}, \mathbf{Q})_{i,j} := (\mathbf{P}\mathbf{e}_j, \mathbf{Q}\mathbf{e}_i)$, where $\mathbf{e}_\ell$ denotes the ℓth column of the $k \times k$ identity matrix.*

In this notation, we have the homogeneity property $(\mathbf{P}\mathbf{L}, \mathbf{Q}) = \mathbf{Q}^*\mathbf{P}\mathbf{L} = (\mathbf{P}, \mathbf{Q})\mathbf{L}$.

Consider the pairs of (time-dependent) functions

$$(\mathbf{Q}, \mathbf{P}), \qquad (\mathbf{P}, \mathbf{Q}) \tag{2.2}$$

and

$$(\mathbf{P}, \mathbf{P}), \qquad (\mathbf{Q}, \mathbf{Q}) \tag{2.3}$$

with derivatives

$$\frac{d}{dt}(\mathbf{Q}, \mathbf{P}) = (\dot{\mathbf{Q}}, \mathbf{P}) + (\mathbf{Q}, \dot{\mathbf{P}}), \quad \frac{d}{dt}(\mathbf{P}, \mathbf{Q}) = (\dot{\mathbf{P}}, \mathbf{Q}) + (\mathbf{P}, \dot{\mathbf{Q}}),$$

and

$$\frac{d}{dt}(\mathbf{P}, \mathbf{P}) = (\dot{\mathbf{P}}, \mathbf{P}) + (\mathbf{P}, \dot{\mathbf{P}}), \quad \frac{d}{dt}(\mathbf{Q}, \mathbf{Q}) = (\dot{\mathbf{Q}}, \mathbf{Q}) + (\mathbf{Q}, \dot{\mathbf{Q}}).$$

Inspired by (1.3), we next investigate how best to choose $\mathbf{L}$ and $\mathbf{M}$ to make either (2.2) or (2.3) invariants of the dynamical system (2.1). In the process we shall see that the pair (2.2) bears a close relationship to coupled "two-sided" iterations, while (2.3) will correspond to "one-sided" iterations.

THEOREM 2.2. *For the system of ordinary differential equations* (2.1) *with initial conditions* $\mathbf{P}(0) = \mathbf{P}_0 \in \mathbb{C}^{n\times k}$ *and* $\mathbf{Q}(0) = \mathbf{Q}_0 \in \mathbb{C}^{n\times k}$*, the choices*

$$\mathbf{L} = (\mathbf{BP}, \mathbf{Q})^{-1}(\mathbf{AP}, \mathbf{Q}), \quad \mathbf{M}^* = (\mathbf{Q}, \mathbf{BP})^{-1}(\mathbf{Q}, \mathbf{AP}).$$

give

$$\frac{d}{dt}(\mathbf{P}, \mathbf{Q}) = \frac{d}{dt}(\mathbf{Q}, \mathbf{P}) = \mathbf{0},$$

and hence $(\mathbf{P}, \mathbf{Q}) = (\mathbf{P}_0, \mathbf{Q}_0)$ *and* $(\mathbf{Q}, \mathbf{P}) = (\mathbf{Q}_0, \mathbf{P}_0)$ *hold for all* t*.*

Proof. Note that

$$\begin{aligned} \frac{d}{dt}(\mathbf{P}, \mathbf{Q}) &= (\dot{\mathbf{P}}, \mathbf{Q}) + (\mathbf{P}, \dot{\mathbf{Q}}) \\ &= (\mathbf{BP}, \mathbf{Q})\mathbf{L} - (\mathbf{AP}, \mathbf{Q}) + \mathbf{M}(\mathbf{P}, \mathbf{B}^*\mathbf{Q}) - (\mathbf{P}, \mathbf{A}^*\mathbf{Q}) \\ \Big(\frac{d}{dt}(\mathbf{Q}, \mathbf{P})\Big)^* &= (\mathbf{P}, \dot{\mathbf{Q}}) + (\dot{\mathbf{P}}, \mathbf{Q}) \\ &= \mathbf{M}(\mathbf{P}, \mathbf{B}^*\mathbf{Q}) - (\mathbf{P}, \mathbf{A}^*\mathbf{Q}) + (\mathbf{BP}, \mathbf{Q})\mathbf{L} - (\mathbf{AP}, \mathbf{Q}), \end{aligned}$$

where we have used (2.1) and the homogeneity property. We can force $(d/dt)(\mathbf{P}, \mathbf{Q})$ to zero by setting

$$\mathbf{L} = (\mathbf{BP}, \mathbf{Q})^{-1}(\mathbf{AP}, \mathbf{Q}), \qquad \mathbf{M} = (\mathbf{P}, \mathbf{A}^*\mathbf{Q})(\mathbf{P}, \mathbf{B}^*\mathbf{Q})^{-1},$$

as given in the theorem. ∎

The next result is a direct analogue of Theorem 2.2 for the pair (2.3). We omit the proof, a minor adaptation of the last one.

THEOREM 2.3. *For the system of ordinary differential equations* (2.1) *with initial conditions* $\mathbf{P}(0) = \mathbf{P}_0 \in \mathbb{C}^{n\times k}$ *and* $\mathbf{Q}(0) = \mathbf{Q}_0 \in \mathbb{C}^{n\times k}$*, the choices*

$$\mathbf{L} = (\mathbf{BP}, \mathbf{P})^{-1}(\mathbf{AP}, \mathbf{P}), \quad \mathbf{M}^* = (\mathbf{Q}, \mathbf{BQ})^{-1}(\mathbf{Q}, \mathbf{AQ})$$

give

$$\frac{d}{dt}(\mathbf{P}, \mathbf{P}) = \frac{d}{dt}(\mathbf{Q}, \mathbf{Q}) = \mathbf{0},$$

and hence $(\mathbf{P}, \mathbf{P}) = (\mathbf{P}_0, \mathbf{P}_0)$ *and* $(\mathbf{Q}, \mathbf{Q}) = (\mathbf{Q}_0, \mathbf{Q}_0)$ *for all* t*.*

With the choices for $\mathbf{L}$ and $\mathbf{M}$ given in Theorems 2.2 and 2.3, we refer to (2.1) as the *two-sided* and *one-sided* dynamical systems.

3. Invariants and backward stability. What do we gain by remaining on the manifold associated with the invariants? The short answer, as we shall see below, is that at all times we can view our state vector as an exact steady-state of a related system. Returning to the subspace setting, if $\dot{\mathbf{P}}$ and $\dot{\mathbf{Q}}$ are small in norm, then that related system can be nearby. Because $\dot{\mathbf{P}}$ and $\dot{\mathbf{Q}}$ are residuals, discussions of nearby systems are the domain of backward stability analysis. See [27] for a comprehensive discussion.

Let us assume the hypothesis of Theorem 2.3. For the one-sided iteration, we rewrite the first part of (2.1) in the form

$$\mathbf{0} = \mathbf{BPL} - (\mathbf{A} + \dot{\mathbf{P}}(\mathbf{P},\mathbf{P})^{-1}\mathbf{P}^*)\mathbf{P}$$

and thus the eigenvalues of $\mathbf{L}$ form a subset of the spectrum of the perturbed pencil $(\mathbf{A}+\dot{\mathbf{P}}(\mathbf{P},\mathbf{P})^{-1}\mathbf{P}^*, \mathbf{B})$. How large can such perturbations be? If $\mathbf{P} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^* \in \mathbb{C}^{n\times k}$ is a singular value decomposition with smallest singular value σ_k, then

$$\|\dot{\mathbf{P}}(\mathbf{P},\mathbf{P})^{-1}\mathbf{P}^*\| \leq \|\dot{\mathbf{P}}\|\|(\mathbf{P},\mathbf{P})^{-1}\mathbf{P}^*\|_2 = \|\dot{\mathbf{P}}\|\,\sigma_k^{-1}.$$

We can thus interpret σ_k^{-1}, which inherits time-invariance from $(\mathbf{P},\mathbf{P})$, as a condition number on the perturbation. Note that if $(\mathbf{P},\mathbf{P}) = \mathbf{I}_k$, then $\sigma_k = 1$ and, provided $\|\dot{\mathbf{P}}\|$ is small, we have the steady-states of a nearby problem. For Hermitian $\mathbf{A}$ and $\mathbf{B} = \mathbf{I}$, Theorem 11.10.1 of Parlett [22] states that the eigenvalue approximations derived from a subspace have condition number σ_k^{-1}, and so basis representation (i.e., departure from orthogonality) matters.

Now consider two-sided iterations, and so assume the hypothesis of Theorem 2.3. We would like to rewrite (2.1) in the form

$$\begin{aligned}\mathbf{0} &= \mathbf{BPL} - (\mathbf{A} + \mathbf{E})\mathbf{P}\\ \mathbf{0} &= \mathbf{B}^*\mathbf{QM}^* - (\mathbf{A}^* + \mathbf{E}^*)\mathbf{Q}\end{aligned}$$

for the same $\mathbf{E}$ in both iterations. If we suppose that $\mathbf{N} = \mathbf{I}$, Lemma 1 of [13] implies that such a perturbation $\mathbf{E}$ exists if and only if

$$(\mathbf{BP},\mathbf{Q})\mathbf{L} = \mathbf{M}(\mathbf{BP},\mathbf{Q}),$$

which holds for the choice of $\mathbf{L}$ and $\mathbf{M}$ given in Theorem 2.2. The perturbation $\mathbf{E}$ is not unique, but $\mathbf{EP} = \dot{\mathbf{P}}$ and $\mathbf{E}^*\mathbf{Q} = \dot{\mathbf{Q}}$. Moreover, the "main theorem" of [13] gives

$$\min \|\mathbf{E}\|_2 = \max\{\|\dot{\mathbf{P}}\|_2, \|\dot{\mathbf{Q}}\|_2\}$$

if $(\mathbf{P},\mathbf{P}) = \mathbf{I}_k$ and $(\mathbf{Q},\mathbf{Q}) = \mathbf{I}_k$. However, as the authors of [13] explain, a small $\|\mathbf{E}\|_2$ is irrelevant unless $\|(\mathbf{P},\mathbf{Q})^{-1}\|_2$ is also small. When $k = 1$, the discussion of Section 4 explains that incurable breakdown occurs when $\mathbf{P}$ is orthogonal to $\mathbf{Q}$, so that $\min \|\mathbf{E}\|_2$ is undefined.

We caution the reader that backward stability alone does not provide information on forward error, or accuracy, of the steady-states when $\mathbf{A} \neq \mathbf{A}^*$. The relevance of backward stability is that the solution of the partitioned ordinary differential equations at all time are steady-states for a related dynamical system. The distance to this related dynamical system depends upon the norm of the residuals.

4. Convergence analysis. At least for single-vector iterations (i.e., $k = 1$), the analysis of the one- and two-sided dynamical systems follows readily from the remarkable fact that, in many cases, simple formulas give the exact solutions of these nonlinear differential equations. This observation, inspired by a lemma of Nanda [19], informs convergence analysis of the eigeniterations that result from the discretization of these equations. Though expressed for the standard eigenvalue problem, these results can naturally be adapted to the generalized case by replacing $\mathbf{A}$ with $\mathbf{B}^{-1}\mathbf{A}$.

THEOREM 4.1. *Consider the partitioned set of ordinary differential equations*

$$\begin{aligned}\dot{\mathbf{p}} &= \mathbf{p}\theta - \mathbf{A}\mathbf{p}\\ \dot{\mathbf{q}} &= \mathbf{q}\bar{\theta} - \mathbf{A}^*\mathbf{q},\end{aligned} \tag{4.1}$$

with $\mathbf{p}(0) = \mathbf{p}_0$ *and* $\mathbf{q}(0) = \mathbf{q}_0$*, where* $\mathbf{p}, \mathbf{q} \in \mathbb{C}^n$*,* $(\mathbf{p}_0, \mathbf{q}_0) \neq 0$*, and*

$$\theta = \frac{(\mathbf{A}\mathbf{p}, \mathbf{q})}{(\mathbf{p}, \mathbf{q})}.$$

Then there exists some $t_{\mathrm{f}} > 0$ *such that for all* $t \in [0, t_{\mathrm{f}})$*,*

$$\mathbf{p}(t) = e^{-\mathbf{A}t}\mathbf{p}_0\pi(t), \quad \mathbf{q}(t) = e^{-\mathbf{A}^*t}\mathbf{q}_0\overline{\pi(t)},$$

where

$$\pi(t) = \sqrt{\frac{(\mathbf{p}_0, \mathbf{q}_0)}{(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0)}}.$$

Proof. We define $\mathbf{p}(t) = e^{-\mathbf{A}t}\mathbf{p}_0\pi(t)$ and $\mathbf{q}(t) = e^{-\mathbf{A}^*t}\mathbf{q}_0\overline{\pi(t)}$, and will show that these formulas satisfy the system (4.1). Note that

$$\begin{aligned}
\dot{\pi} &= \frac{\pi}{2}\frac{\Big((\mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0) + (e^{-\mathbf{A}t}\mathbf{p}_0, \mathbf{A}^*e^{-\mathbf{A}^*t}\mathbf{q}_0)\Big)}{(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0)} \\
&= \pi\frac{(\mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0)}{(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0)} \\
&= \pi\frac{(\mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0\pi, e^{-\mathbf{A}^*t}\mathbf{q}_0\bar{\pi})}{(e^{-\mathbf{A}t}\mathbf{p}_0\pi, e^{-\mathbf{A}^*t}\mathbf{q}_0\bar{\pi})} \;=\; \pi\frac{(\mathbf{A}\mathbf{p}, \mathbf{q})}{(\mathbf{p}, \mathbf{q})} \;=\; \pi\theta.
\end{aligned}$$

Differentiating the formulas for $\mathbf{p}$ and $\mathbf{q}$ thus gives

$$\begin{aligned}
\dot{\mathbf{p}} &= -\mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0\pi + e^{-\mathbf{A}t}\mathbf{p}_0\dot{\pi} \;=\; -\mathbf{A}\mathbf{p} + \theta\mathbf{p} \\
\dot{\mathbf{q}} &= -\mathbf{A}^*e^{-\mathbf{A}^*t}\mathbf{q}_0\bar{\pi} + e^{-\mathbf{A}^*t}\mathbf{q}_0\dot{\bar{\pi}} \;=\; -\mathbf{A}^*\mathbf{q} + \bar{\theta}\mathbf{q},
\end{aligned}$$

as required. The hypothesis that $(\mathbf{p}_0, \mathbf{q}_0) \neq 0$ ensures the existence of the solution at time $t = 0$. The formula will hold for all $t > 0$, until potentially

$$(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0) = 0. \tag{4.2}$$

We define t_{f} to be the smallest positive t for which (4.2) holds. If no such positive t exists, the solution exists for all $t > 0$ and we can take $t_{\mathrm{f}} = \infty$ in the statement of the theorem. ∎

Theorem 4.1 gives $(\mathbf{p}, \mathbf{q}) = (\mathbf{p}_0, \mathbf{q}_0)$, precisely as Theorem 2.2 indicates. Under the conditions of Theorem 4.1, solutions of the two-sided single-vector equations (4.1) have the same direction as solutions of the simpler linear systems $\dot{\mathbf{x}} = -\mathbf{A}\mathbf{x}$, $\mathbf{x}(0) = \mathbf{p}_0$ and $\dot{\mathbf{y}} = -\mathbf{A}^*\mathbf{y}$, $\mathbf{y}(0) = \mathbf{q}_0$, but the magnitudes of $\mathbf{p}$ and $\mathbf{q}$ in (4.1) vary nonlinearly. In particular, this magnitude can blow-up in finite time—a phenomenon we call *incurable breakdown*—even with both $\mathbf{p}$ and $\mathbf{q}$ nonzero. Note that if

$$\left(\frac{e^{-\mathbf{A}t}\mathbf{p}_0}{\sqrt{(\mathbf{p}_0, \mathbf{q}_0)}}, \frac{e^{-\mathbf{A}^*t}\mathbf{q}_0}{\sqrt{(\mathbf{q}_0, \mathbf{p}_0)}}\right) = 0$$

then $\pi(t)$ is undefined. This ratio will be nonzero but small in the vicinity of blow-up, a situation that commonly occurs in discretizations of these equations. The salient issue is that $\mathbf{p}$ and $\mathbf{q}$ are nearly orthogonal and so

$$\frac{(\mathbf{p}, \mathbf{q})}{\|\mathbf{p}\|\,\|\mathbf{q}\|} = \left(\frac{e^{-\mathbf{A}t}\mathbf{p}_0}{\|e^{-\mathbf{A}t}\mathbf{p}_0\|}, \frac{e^{-\mathbf{A}^*t}\mathbf{q}_0}{\|e^{-\mathbf{A}^*t}\mathbf{q}_0\|}\right) \tag{4.3}$$

is a useful quantity to measure. This number is small when the secant of the angle between $\mathbf{p}$ and $\mathbf{q}$ is large. In Section 6 we shall see the important consequences of these observations for eigensolvers derived from the discretization of (4.1).

One can avoid breakdown altogether by using starting vectors $\mathbf{p}_0$ and $\mathbf{q}_0$ that are sufficiently accurate approximations to the right and left eigenvectors of $\mathbf{A}$ associated with the leftmost eigenvalue. Suppose $\mathbf{A}$ is diagonalizable with a simple leftmost eigenvalue λ_1, and all other eigenvalues strictly to the right of λ_1. Thus there exists invertible $\mathbf{X}$ and diagonal $\mathbf{\Lambda}$ such that

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$$

with $\mathbf{\Lambda}_{1,1} = \lambda_1$. Write $\lambda_j = \mathbf{\Lambda}_{j,j}$, so that $\operatorname{Re}\lambda_j > \operatorname{Re}\lambda_1$ for $j = 2, \ldots, n$. Define $\mathbf{r} = \mathbf{X}^{-1}\mathbf{p}_0$ and $\mathbf{s} = \mathbf{X}^*\mathbf{q}_0$, i.e., $\mathbf{r}$ and $\mathbf{s}$ are the expansions of the starting vectors in biorthogonal bases of right and left eigenvectors of $\mathbf{A}$.

THEOREM 4.2. *Under the setting established in the last paragraph, the condition*

$$|r_1 s_1| > \sum_{j=2}^{n} |r_j s_j|$$

is sufficient to ensure that the dynamical system (4.1) *has a solution for all* $t \geq 0$ *given by Theorem 4.1, i.e., no incurable breakdown occurs.*

Proof. First note that

$$(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0) = (\mathbf{X}e^{-\mathbf{\Lambda}}\mathbf{X}^{-1}\mathbf{p}_0, \mathbf{X}^{-*}e^{-\mathbf{\Lambda}^*}\mathbf{X}^*\mathbf{q}_0) = (e^{-2\mathbf{\Lambda}t}\mathbf{r}, \mathbf{s}) = \sum_{j=1}^{n} r_j \overline{s}_j e^{-2\lambda_j t}.$$

Since $\operatorname{Re}\lambda_1 < \operatorname{Re}\lambda_j$ for $j > 2$, we have $|e^{-2\lambda_1 t}| \geq |e^{-2\lambda_j t}|$ for all $t \geq 0$. The hypothesis involving $\mathbf{r}$ and $\mathbf{s}$ thus implies, for $t \geq 0$, that

$$|r_1 s_1 e^{-2\lambda_1 t}| \geq \sum_{j=2}^{n} |r_j s_j e^{-2\lambda_j t}|.$$

Given this expression, we can twice apply the triangle inequality to conclude

$$\begin{aligned} 0 &< |r_1 \overline{s}_1 e^{-2\lambda_1 t}| - \sum_{j=2}^{n} |r_j \overline{s}_j e^{-2\lambda_j t}| \\ &\leq |r_1 \overline{s}_1 e^{-2\lambda_1 t}| - \left| \sum_{j=2}^{n} r_j \overline{s}_j e^{-2\lambda_j t} \right| \leq \left| \sum_{j=1}^{n} r_j \overline{s}_j e^{-2\lambda_j t} \right| = |(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0)|. \end{aligned}$$

Hence $\pi(t)$ in Theorem 4.1 is finite for all $t \geq 0$, ensuring that the solution to the dynamical system (4.1) does not blow up at finite time. ∎

The single vector one-sided system possesses a similar exact solution, which has been studied in the context of gradient flows associated with Rayleigh quotient iteration. We shall see that finite-time blow-up is never a concern for such systems. The following is a modest restatement of a result of Nanda [19, Lemma 1.4] (who considers the differential equation acting on the unit ball in $\mathbb{R}^n$).

THEOREM 4.3. *Consider the ordinary differential equation*

$$\dot{\mathbf{p}} = \mathbf{p}\theta - \mathbf{A}\mathbf{p}, \tag{4.4}$$

with $\mathbf{A} \in \mathbb{R}^{n\times n}$ *and initial condition* $\mathbf{p}(0) = \mathbf{p}_0 \in \mathbb{R}^n$*, where* $\mathbf{p}_0 \neq \mathbf{0}$ *and*

$$\theta = \frac{(\mathbf{A}\mathbf{p}, \mathbf{p})}{(\mathbf{p}, \mathbf{p})}.$$

Then for all $t \geq 0$*, equation* (4.4) *has the exact solution*

$$\mathbf{p}(t) = e^{-\mathbf{A}t}\mathbf{p}_0\omega(t)$$

where

$$\omega(t) = \sqrt{\frac{(\mathbf{p}_0, \mathbf{p}_0)}{(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}t}\mathbf{p}_0)}}.$$

We omit the proof of this result, which closely mimics that of Theorem 4.1. Of course, a similar formula can be written for the one-sided equation for $\mathbf{q}(t)$. The restriction to real matrices guarantees that $(\mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}t}\mathbf{p}_0) = (e^{-\mathbf{A}t}\mathbf{p}_0, \mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0)$; the result also hold for complex Hermitian $\mathbf{A}$.

As before, $\mathbf{p}$ has the same direction as the solution to the dynamical system $\dot{\mathbf{x}} = -\mathbf{A}\mathbf{x}$ with $\mathbf{x}(0) = \mathbf{p}_0$, but the magnitude is scaled by the nonlinear scalar ω. Provided $\mathbf{p}_0 \neq \mathbf{0}$, the one-sided system (4.4) cannot blow up in finite time, since $(\mathbf{p}, \mathbf{p}) \neq 0$, in stark contrast to the two-sided iteration. This collinearity implies that the $\mathbf{p}$ vectors produced by the one- and two-sided systems provide equally accurate approximations to the desired eigenvector, at least until the latter breaks down.

When $\mathbf{A}$ has a unique simple eigenvalue of smallest real part and the hypotheses of Theorem 4.1 or 4.3 are met, the asymptotic analysis of the associated dynamical system readily follows; cf. [12, §1.3] for a generic asymptotic linear stability analysis of the one-sided iteration. In fact, one can develop explicit bounds on the sine of the angle between $\mathbf{p}$ and the desired eigenvector $\mathbf{x}_1$, defined as

$$\sin\angle(\mathbf{p}, \mathbf{x}_1) := \min_{\alpha\in\mathbb{C}} \frac{\|\alpha\mathbf{p} - \mathbf{x}_1\|}{\|\mathbf{x}_1\|}.$$

THEOREM 4.4. *Suppose* $\mathbf{A}$ *can be diagonalized,* $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$*, and the eigenvalues of* $\mathbf{A}$ *can be ordered as*

$$\mathrm{Real}(\lambda_1) < \mathrm{Real}(\lambda_2) \leq \cdots \leq \mathrm{Real}(\lambda_n).$$

Let $\mathbf{x}_1$ *and* $\mathbf{y}_1$ *denote right and left eigenvectors associated with* λ_1*, with* $\|\mathbf{x}_1\| = 1$ *and* $\mathbf{y}_1^*\mathbf{x}_1 = 1$*. Then the solution* $\mathbf{p}(t)$ *to both systems* (4.1) *and* (4.4) *satisfies*

$$\sin\angle(\mathbf{p}(t), \mathbf{x}_1) \leq \|\mathbf{X}\|\,\|\mathbf{X}^{-1}\|\,\frac{\|\mathbf{p}_0\|}{|\mathbf{y}_1^*\mathbf{p}_0|}e^{\mathrm{Re}(\lambda_1-\lambda_2)t}$$

for all $t \geq 0$ *in the case of* (4.4)*, and for all* $t \in [0, t_f)$ *in the case of* (4.1)*.*

Proof. Since $\mathbf{x}_1$ is a unit vector, we can write

$$\sin\angle(\mathbf{p}(t), \mathbf{x}_1) = \min_{\alpha\in\mathbb{C}} \|\alpha\mathbf{p}(t) - \mathbf{x}_1\|.$$

In both (4.4) and (4.1), $\mathbf{p}(t)$ is collinear with $e^{-\mathbf{A}t}\mathbf{p}_0$, so we can proceed with

$$\begin{aligned}\sin\angle(\mathbf{p}(t), \mathbf{x}_1) &= \min_{\alpha\in\mathbb{C}} \|\alpha\mathbf{X}e^{-\mathbf{\Lambda}t}\mathbf{X}^{-1}\mathbf{p}_0 - \mathbf{x}_1\| \\ &\leq \left\|\frac{e^{\lambda_1 t}}{\mathbf{y}_1^*\mathbf{p}_0}\mathbf{X}e^{-\mathbf{\Lambda}t}\mathbf{X}^{-1}\mathbf{p}_0 - \mathbf{x}_1\right\| \;\leq\; \|\mathbf{X}\|\,\|\mathbf{X}^{-1}\|\,\frac{\|\mathbf{p}_0\|}{|\mathbf{y}_1^*\mathbf{p}_0|}e^{\mathrm{Re}(\lambda_1-\lambda_2)t}.\end{aligned}$$

The first inequality follows from choosing a (sub-optimal) value of α that cancels the terms in the $\mathbf{x}_1$ direction. (For similar analysis of the Arnoldi eigenvalue iteration, see [26, Prop. 2.1].) □

An analogous bound could be developed for the convergence of $\mathbf{q}$ to the left eigenvector $\mathbf{y}_1$. When $\mathbf{A}$ is far from normal, one typically observes a transient stage of convergence that could be better described via analysis that avoids the diagonalization of $\mathbf{A}$; see, e.g.,[29, §28], which includes similar analysis for the power method.

The two-sided iteration converges to left and right eigenvectors of $\mathbf{A}$ associated with the leftmost eigenvalue, *provided the method does not breakdown on the way to this limit*. Several natural questions arise: How common is breakdown? How well do discretizations mimic this dynamical system? Before investigating these issues in Section 6, we first address how preconditioning can accelerate—and complicate—the convergence of these continuous-time systems.

5. Preconditioned dynamical systems. What does it mean to precondition the eigenvalue problem? Several different strategies have been proposed in the literature (see especially the discussion in [14, pp. 109–110]); here we shall investigate analogous approaches for our continuous time dynamical systems, and the implications such modifications have on the convergence behavior described in the last section.

One might first consider applying to the generalized eigenvalue problem

$$\mathbf{A}\mathbf{p} = \mathbf{B}\mathbf{p}\lambda,$$

left and right preconditioners $\mathbf{M}$ and $\mathbf{N}$, so as to obtain the equivalent pencil

$$(\mathbf{M}^{-1}\mathbf{A}\mathbf{N})(\mathbf{N}^{-1}\mathbf{p}) = (\mathbf{M}^{-1}\mathbf{B}\mathbf{N})(\mathbf{N}^{-1}\mathbf{p})\lambda. \tag{5.1}$$

Provided $\mathbf{B}$ is invertible, one could then define

$$\widehat{\mathbf{A}} := (\mathbf{M}^{-1}\mathbf{B}\mathbf{N})^{-1}(\mathbf{M}^{-1}\mathbf{A}\mathbf{N}) = \mathbf{N}^{-1}\mathbf{B}^{-1}\mathbf{A}\mathbf{N}$$

$$\widehat{\mathbf{p}} := \mathbf{N}^{-1}\mathbf{p},$$

then apply the concepts from the preceding sections to the standard eigenvalue problem $\widehat{\mathbf{A}}\widehat{\mathbf{p}} = \widehat{\mathbf{p}}\lambda$. For example, we could seek the leftmost eigenpair of $\widehat{\mathbf{A}}$ by evolving the dynamical system

$$\dot{\widehat{\mathbf{p}}} = \widehat{\mathbf{p}}\widehat{\theta} - \widehat{\mathbf{A}}\widehat{\mathbf{p}},$$

with the (preconditioned) Rayleigh quotient

$$\widehat{\theta} = \frac{(\widehat{\mathbf{A}}\widehat{\mathbf{p}}, \widehat{\mathbf{p}})}{(\widehat{\mathbf{p}}, \widehat{\mathbf{p}})} = \frac{(\mathbf{N}^{-1}\mathbf{B}^{-1}\mathbf{A}\mathbf{p}, \mathbf{N}^{-1}\mathbf{p})}{(\mathbf{N}^{-1}\mathbf{p}, \mathbf{N}^{-1}\mathbf{p})}.$$

Note that $\widehat{\mathbf{A}}$ and $\mathbf{B}^{-1}\mathbf{A}$ share the same spectrum since they are similar, and hence the asymptotic rate in Theorem 4.4 is immune to the preconditioner. The application of $\mathbf{N}$ could affect the system's transient behavior, but $\mathbf{M}$ exerts no influence at all.[1]

Several choices for $\mathbf{N}$ are interesting. Taking $\mathbf{N} = \mathbf{A}^{-1}$ gives $\widehat{\mathbf{A}} = \mathbf{A}\mathbf{B}^{-1}$, an alternative to the $\mathbf{B}^{-1}\mathbf{A}$ form suggested by the original problem. Similarity transformations can also be used to *balance* a matrix to improve the conditioning of the

[1] Alternatively, by substituting $(\mathbf{M}^{-1}\mathbf{B}\mathbf{N})^{-1}\widetilde{\mathbf{p}} := \mathbf{N}^{-1}\mathbf{p}$ in equation (5.1), we obtain a system driven by $\widetilde{\mathbf{A}} = \mathbf{M}^{-1}\mathbf{A}\mathbf{B}^{-1}\mathbf{M}$ that is independent of $\mathbf{N}$.

eigenvalue problem [21, 23], in which case $\mathbf{N}$ is constructed as a diagonal matrix that reduces the norm of $\widehat{\mathbf{A}}$. Such balancing tends to decrease the departure from normality associated with the largest magnitude eigenvalues. In fact, in the 1960 article that introduced this idea, Osborne refers to this procedure as "pre-conditioning" [21]. A more extreme—if impractical—approach takes $\mathbf{N}$ to be a matrix that diagonalizes $\mathbf{B}^{-1}\mathbf{A}$ (provided such a matrix exists), a choice that minimizes the constant $\|\mathbf{X}\|\|\mathbf{X}^{-1}\|$ that describes the departure from normality in Theorem 4.4.

As useful as such improvements might be, these strategies fail to alter the asymptotic convergence rate described in Theorem 4.4. To potentially improve this rate, one can apply the preconditioner $\mathbf{N}^{-1}$ directly to the residual $\mathbf{p}\theta - \mathbf{A}\mathbf{p}$. Consider the dynamical system

$$\dot{\mathbf{p}} = \mathbf{N}^{-1}(\mathbf{p}\theta - \mathbf{A}\mathbf{p}), \tag{5.2}$$

where θ refers to the usual (unpreconditioned) Rayleigh quotient $\theta = (\mathbf{A}\mathbf{p}, \mathbf{p})/(\mathbf{p}, \mathbf{p})$. Discretization of this system results in the familiar preconditioned eigensolver described in (1.1). For this case, a generalization of Theorem 4.3 has proved elusive; we have found no closed form for the exact solution. Indeed, as we shall next see, the choice of preconditioner can even complicate the system's local behavior.

Let $\mathbf{x}_1$ denote a unit eigenvector of $\mathbf{A}$ associated with the eigenvector λ_1, which we assume to be strictly to the left of the other eigenvalues of $\mathbf{A}$. Note that $\mathbf{x}_1$ is a steady-state of (5.2), linearizing about which gives the Jacobian

$$\mathbf{J} = \mathbf{N}^{-1}(\mathbf{I} - \mathbf{x}_1\mathbf{x}_1^*)(\lambda_1 - \mathbf{A}). \tag{5.3}$$

As $\mathbf{J}\mathbf{x}_1 = \mathbf{0}$, the Jacobian $\mathbf{J}$ always has a zero eigenvalue, adding complexity to conventional linear stability analysis. The challenge can be magnified by a poor choice for $\mathbf{N}$. For example, suppose

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad \mathbf{N} = \mathbf{N}^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \lambda_1 = 1,$$

so that

$$\mathbf{J} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix},$$

i.e., the Jacobian is a Jordan block with a double eigenvalue at zero.

To obtain a rough impression of the behavior of the continuous system when θ is in the vicinity of λ_1, consider the constant-coefficient equation $\dot{\mathbf{p}} = \mathbf{N}^{-1}(\mathbf{p}\lambda_1 - \mathbf{A}\mathbf{p})$, whose solution obeys the simple formula

$$\mathbf{p}(t) = e^{\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})t}\mathbf{p}(0).$$

Hence the asymptotic behavior of $\mathbf{p}$ is controlled by the spectrum of $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$. Assuming that $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$ has a simple zero eigenvalue, the convergence of this system to the dominant eigenvector depends on the nonzero eigenvalues of $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$: if this matrix has any other eigenvalues in the closed right half plane, the system will not generically converge; if all nonzero eigenvalues are in the open left half plane, then the convergence rate will be determined by the rightmost of them.

From the perspective of the convergence rate of the continuous dynamical system, we seek a preconditioner $\mathbf{N}^{-1}$ such that the nonzero eigenvalues of $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$ are

as far to the left as possible. While the leftmost eigenvalues of $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$ do not much affect the behavior of the continuous system, they can have a significant effect on the stability of the discretized difference equation, i.e., the related eigensolvers. For example, if $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$ moves all nonzero eigenvalues into the left half plane, then replacing $\mathbf{N}$ by $\frac{1}{2}\mathbf{N}$ doubles the convergence rate of the continuous system.

To rigorously analyze the local behavior of the fully nonlinear system when $\mathbf{p}$ approximates the eigenvector $\mathbf{x}_1$, we shall apply the Center Manifold Theorem [5, 10], a tool for studying a dynamical system whose Jacobian has an eigenvalue on the imaginary axis. (Alternatively, we could restrict the system to the unit sphere in $\mathbb{R}^n$.) Without loss of generality, assume that $\lambda_1 = 0$, so that the Jacobian at $\mathbf{x}_1$ (5.3) takes the form $\mathbf{J} = -\mathbf{N}^{-1}(\mathbf{I} - \mathbf{x}_1\mathbf{x}_1^*)\mathbf{A}$. Thus for $\mathbf{p}$ near $\mathbf{x}_1$ we have

$$\dot{\mathbf{p}} = \mathbf{J}\mathbf{p} + \mathbf{F}(\mathbf{p})$$

for the nonlinear function $\mathbf{F}(\mathbf{p}) = \mathbf{N}^{-1}(\theta(\mathbf{p})\mathbf{p} - (\mathbf{A}\mathbf{p}, \mathbf{x}_1)\mathbf{x}_1)$ that, by definition of the Jacobian, satisfies $\|\mathbf{F}(\mathbf{p})\| = o(\|\mathbf{p} - \mathbf{x}_1\|)$.

Suppose that $\mathbf{J}$ has a simple zero eigenvalue, and the rest of its spectrum is in the open left half plane. There exists some invertible (real, if $\mathbf{J}$ is real) matrix $\mathbf{S}$ with first column $\mathbf{x}_1$ and

$$\mathbf{S}^{-1}\mathbf{J}\mathbf{S} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix}$$

for some $\mathbf{C} \in \mathbb{C}^{(n-1)\times(n-1)}$ whose spectrum is in the open left half plane.

We now transform coordinates into a form in which the Center Manifold Theorem can most readily be applied. Define

$$\mathbf{r}(t) = \mathbf{S}^{-1}(\mathbf{p}(t) - \mathbf{x}_1),$$

so that

$$\dot{\mathbf{r}} \;=\; (\mathbf{S}^{-1}\mathbf{J}\mathbf{S})\mathbf{S}^{-1}(\mathbf{p} - \mathbf{x}_1) + \mathbf{S}^{-1}\mathbf{F}(\mathbf{p}) \;=\; \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix}\mathbf{r} + \mathbf{G}(\mathbf{r}),$$

where $\mathbf{G}(\mathbf{r}) := \mathbf{S}^{-1}\mathbf{F}(\mathbf{S}\mathbf{r} + \mathbf{x}_1) = \mathbf{S}^{-1}\mathbf{F}(\mathbf{p})$. By design, $\mathbf{S}^{-1}\mathbf{x}_1 = \mathbf{e}_1$, and hence $\mathbf{G}(\mathbf{r})$ satisfies

$$\mathbf{G}(\mathbf{r}) = \mathbf{S}^{-1}\mathbf{N}^{-1}\mathbf{S}\Big(\Big(\frac{(\mathbf{A}\mathbf{S}\mathbf{r}, \mathbf{S}\mathbf{r}) + (\mathbf{A}\mathbf{S}\mathbf{r}, \mathbf{x}_1)}{(\mathbf{S}\mathbf{r}, \mathbf{S}\mathbf{r}) + 2(\mathbf{x}_1, \mathbf{S}\mathbf{r}) + 1} \Big)(\mathbf{r} + \mathbf{e}_1) - ((\mathbf{A}\mathbf{S}\mathbf{r}, \mathbf{x}_1)\mathbf{e}_1 \Big). \qquad (5.4)$$

Now we are prepared to cast this diagonalized problem into the conventional setting for Center Manifold Theory. We write

$$\mathbf{r} = \begin{bmatrix} \alpha \\ \mathbf{b} \end{bmatrix}$$

for $\alpha \in \mathbb{R}$ and $\mathbf{b} \in \mathbb{R}^{n-1}$. Using MATLAB index notation for convenience, the $\mathbf{r}$ system is simply

$$\begin{bmatrix} \dot{\alpha} \\ \dot{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix} \begin{bmatrix} \alpha \\ \mathbf{b} \end{bmatrix} + \begin{bmatrix} \mathbf{G}([\alpha; \mathbf{b}])_1 \\ \mathbf{G}([\alpha; \mathbf{b}])_{2:n} \end{bmatrix}$$

that is

$$\dot{\alpha} = \mathbf{G}([\alpha; \mathbf{b}])_1, \quad \dot{\mathbf{b}} = \mathbf{C}\mathbf{b} + \mathbf{G}([\alpha; \mathbf{b}])_{2:n}.$$

Notice that the component α only figures in the nonlinear terms; we wish to determine how that contribution affects the magnitude of the $\mathbf{b}$ component—that is, the portion of the solution that we hope decays as $t \to \infty$. Notice that $\mathbf{b} = \mathbf{0}$ corresponds to the case when $\mathbf{p}$ is collinear with $\mathbf{x}_1$. In this case $\mathbf{p}$ may differ from the unit eigenvector $\mathbf{x}_1$, but regardless it is a fixed point of the dynamical system, and provided $\mathbf{p} \neq \mathbf{0}$ we are content. In particular, if $\mathbf{b} = \mathbf{0}$, then $\mathbf{ASr} = \mathbf{0}$ too (recall that $\lambda = 0$), and we can see from (5.4) that $\mathbf{G}(\mathbf{r}) = \mathbf{0}$. In this case

$$\dot{\alpha} = \mathbf{G}([\alpha; \mathbf{0}])_1 = 0, \quad \dot{\mathbf{b}} = \mathbf{C0} + \mathbf{G}([\alpha; \mathbf{0}])_{2:n} = \mathbf{0},$$

so any such $\mathbf{r}$ is a fixed point of the dynamical system. We can put this in grander language: there exists some $\delta > 0$ such that if

$$\mathbf{r}_0 \in \left\{ \begin{bmatrix} \alpha \\ \mathbf{0} \end{bmatrix} : |x| < \delta \right\} =: \mathcal{M},$$

then the dynamical system with $\mathbf{r}(0) = \mathbf{r}_0$ satisfies $\mathbf{r}(t) \in \mathcal{M}$ for all $t > 0$. (In particular, $\mathbf{r}(t) = \mathbf{r}(0) \in \mathcal{M}$.) The set $\mathcal{M}$ is called a *local invariant manifold*. We can define this manifold (locally) by the requirement that

$$\mathbf{b} = \mathbf{g}(\alpha) := \mathbf{0},$$

which trivially satisfies $\mathbf{g}(0) = \mathbf{0}$ and the Jacobian of $\mathbf{g}$ at $\alpha = 0$ is $D\mathbf{g}(0) = \mathbf{0}$; furthermore, $\mathbf{g}$ is arbitrarily smooth near $\alpha = 0$. Together, these properties ensure that $\mathcal{M}$ is a *center manifold* of the dynamical system. (We are fortunate in this case to have an explicit, trivial expression for this manifold.)

All that remains is to apply Theorem 2 from Carr [5, p. 4]. Consider the equation

$$\dot{u} = \mathbf{G}([u; \mathbf{g}(u)])_1 = \mathbf{G}([u; \mathbf{0}])_1 = 0.$$

The solution $u(t) = 0$ is clearly stable—if $u(t) = \varepsilon$, then $|u(t) - 0| = |\varepsilon|$ is bounded for all $t > 0$—and thus Theorem 2(a) from [5] implies that the solution $\mathbf{r}(t) = \mathbf{0}$ is a stable solution of the system

$$\dot{\mathbf{r}} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix} \mathbf{r} + \mathbf{G}(\mathbf{r}).$$

Note that the solution $u(t) = 0$ is not *asymptotically stable*, that is, we do not have $u(t) \to 0$ if $u(0) = \varepsilon$ for small, nonzero ε. Were this the case, then we would be able to conclude that the $\mathbf{r}$ system was asymptotically stable. This would contradict our expectation that the original dynamical system will converge to something in $\operatorname{span}\{\mathbf{x}_1\}$, not necessarily to $\mathbf{x}_1$ itself. In particular, if $\mathbf{N}$ is self-adjoint, then $(\mathbf{Np}, \mathbf{p})$ is an invariant of the system, and so we expect that $\mathbf{p}(t) \to \xi \mathbf{x}_1$ for ξ determined by

$$|\xi|^2 = \frac{(\mathbf{Np}, \mathbf{p})}{(\mathbf{Nx}_1, \mathbf{x}_1)}.$$

We now have stability of the zero state of the $\mathbf{r}$ system, but that only means that solutions sufficiently close to $\mathbf{r} = \mathbf{0}$ do not diverge. To say more—to say that the solutions actually converge to the center manifold—we can apply Theorem 2(b) of [5], which we slightly paraphrase here. Since the zero solution of the $\mathbf{r}$ equation

is stable, for $\|[\alpha(0); \mathbf{b}(0)]\|$ sufficiently small, there exists some solution $u(t)$ of the equation $\dot{u}(t) = \mathbf{G}([u; \mathbf{g}(u)])_1 = 0$ and positive constant γ such that

$$\alpha(t) = u(t) + O(e^{-\gamma t}), \quad \mathbf{b}(t) = \mathbf{g}(u(t)) + O(e^{-\gamma t}).$$

In particular, in our setting such solutions $u(t)$ will be constant: $u(t) = c$, and so there exist

$$\alpha(t) = c + O(e^{-\gamma t}), \quad \mathbf{b}(t) = O(e^{-\gamma t}),$$

and in particular $\|\mathbf{b}(t)\| \to 0$ as $t \to \infty$. Thus for $\|\mathbf{r}_0\|$ sufficiently small,

$$\mathbf{r}(t) = \begin{bmatrix} c \\ \mathbf{0} \end{bmatrix} + O(e^{-\gamma t}),$$

so that $\mathbf{p}(t) = \mathbf{S}\mathbf{r}(t) + \mathbf{x}_1 = (1 + c)\mathbf{x}_1 + O(e^{-\gamma t})$. The preceding discussion is summarized in the following result.

THEOREM 5.1. *If $\|\mathbf{p}(0) - \mathbf{x}_1\|$ is sufficiently small and $\mathbf{N}^{-1}(\mathbf{I} - \mathbf{x}_1\mathbf{x}_1^*)(\lambda - \mathbf{A})$ has a simple zero eigenvalue with all other eigenvalues in the open left half plane, then there exists $\gamma > 0$ and $\xi \in \mathbb{R}$ such that, as $t \to \infty$,*

$$\|\mathbf{p}(t) - \xi\mathbf{x}_1\| = O(e^{-\gamma t}).$$

In the case of self-adjoint, invertible $\mathbf{N}$, $|\xi| = |(\mathbf{p}_0, \mathbf{N}\mathbf{p}_0)|$.

Note that if $\mathbf{N}$ is Hermitian and invertible but indefinite, then there always exists some unit vector $\mathbf{p}_0$ such that $(\mathbf{p}_0, \mathbf{N}\mathbf{p}_0) = 0$. If this starting vector is sufficiently close to the unit eigenvector $\mathbf{x}_1$ of $\mathbf{A}$, then we have not ruled out the possibility that the system converges to the zero vector, rather than a desired eigenvector.

6. Discrete dynamical systems. The previous sections have addressed the quadratic invariant and convergence behavior of the continuous-time one- and two-sided dynamical systems. For purposes of computation, one naturally wonders how closely such properties are mimicked by the solutions to discretizations of these systems. The present section considers the convergence and preservation of the quadratic invariant by the discrete flow under a forward Euler time integration.

6.1. Departure from the manifold. Given $\mathbf{A} \in \mathbb{R}^{n \times n}$, for notational convenience we rewrite the two-sided system in the form

$$\begin{array}{rcll} \dot{\mathbf{p}} & = & \mathbf{p}\theta - \mathbf{A}\mathbf{p} & =: \mathbf{f}(\mathbf{p}, \mathbf{q}) \\ \dot{\mathbf{q}} & = & \mathbf{q}\theta - \mathbf{A}^T\mathbf{q} & =: \mathbf{g}(\mathbf{p}, \mathbf{q}), \end{array} \tag{6.1}$$

with $\theta = (\mathbf{q}^T\mathbf{p})^{-1}\mathbf{q}^T\mathbf{A}\mathbf{p} = \theta^T$ and initial conditions $\mathbf{p}(0) = \mathbf{p}_0 \in \mathbb{R}^n$ and $\mathbf{q}(0) = \mathbf{q}_0 \in \mathbb{R}^n$. Similarly, the one-sided system (now including preconditioning) is

$$\dot{\mathbf{p}} = \mathbf{N}^{-1}(\mathbf{p}\theta - \mathbf{A}\mathbf{p}) =: \mathbf{N}^{-1}\mathbf{f}(\mathbf{p}, \mathbf{p}). \tag{6.2}$$

with $\theta = (\mathbf{p}^T\mathbf{p})^{-1}\mathbf{p}^T\mathbf{A}\mathbf{p} = \theta^T$ and $\mathbf{p}(0) = \mathbf{p}_0 \in \mathbb{R}^n$.

In Section 2 we showed that this system preserves the quadratic invariant $\mathbf{q}^T\mathbf{p}$. To what extent do discretizations respect such conservation, and what are the implications of any drift from this manifold?

In order to understand the role of discrete quadratic invariants, we consider the error when using a forward Euler time integrator on the one- and two-sided iterations. The resulting discretization of (6.1) leads to the discrete dynamical system

$$\mathbf{p}_{j+1} = \mathbf{p}_j + h\mathbf{f}_j \tag{6.3}$$
$$\mathbf{q}_{j+1} = \mathbf{q}_j + h\mathbf{g}_j, \tag{6.4}$$

where $\mathbf{f}_j := \mathbf{f}(\mathbf{p}_j, \mathbf{q}_j)$ and $\mathbf{g}_j := \mathbf{g}(\mathbf{p}_j, \mathbf{q}_j)$. With the mild caveat that $\mathbf{q}_j^T \mathbf{p}_j \neq 0$, the form of the Rayleigh quotient gives

$$\mathbf{q}_j^T \mathbf{f}_j = 0 = \mathbf{p}_j^T \mathbf{g}_j.$$

This simple observation is critical to understanding the drift of the forward Euler iterates from the invariant manifold. It implies, for example, that the first iteration of (6.3)–(6.4) produces a iterate that is quadratically close to the manifold:

$$\mathbf{q}_1^T \mathbf{p}_1 = \mathbf{q}_0^T \mathbf{p}_0 + h^2(\mathbf{g}_0^T \mathbf{f}_0),$$

which is perhaps surprising given the forward Euler method's $O(h)$ accuracy. Writing the departure from the manifold as

$$d_j = \mathbf{q}_j^T \mathbf{p}_j - \mathbf{q}_0^T \mathbf{p}_0,$$

we thus have $d_1 = h^2(\mathbf{g}_0^T \mathbf{f}_0)$. From this we can compute

$$d_2 = (\mathbf{q}_2^T \mathbf{p}_2 - \mathbf{q}_1^T \mathbf{p}_1) + d_1 = h^2(\mathbf{g}_1^T \mathbf{f}_1 + \mathbf{g}_0^T \mathbf{f}_0)$$

and, in general, $d_{j+1} = h^2 \sum_{k=0}^{j} \mathbf{g}_k^T \mathbf{f}_k$. (This result is a special case of one derived in [11] for partitioned Runge–Kutta systems.) Thus we can bound the relative drift from the manifold as

$$\frac{|\mathbf{q}_{j+1}^T \mathbf{p}_{j+1} - \mathbf{q}_0^T \mathbf{p}_0|}{|\mathbf{q}_0^T \mathbf{p}_0|} \leq h^2 \sum_{k=0}^{j} \frac{\|\mathbf{f}_k\| \, \|\mathbf{g}_k\|}{|\mathbf{q}_0^T \mathbf{p}_0|}. \tag{6.5}$$

The definitions of $\mathbf{f}(\mathbf{p}, \mathbf{q})$ and $\mathbf{g}(\mathbf{p}, \mathbf{q})$ imply

$$\|\mathbf{f}_k\| \leq (|\theta_k| + \|\mathbf{A}\|) \, \|\mathbf{p}_k\| \leq \left(1 + \frac{\|\mathbf{q}_k\| \, \|\mathbf{p}_k\|}{|\mathbf{q}_k^T \mathbf{p}_k|}\right) \|\mathbf{A}\| \, \|\mathbf{p}_k\|$$

$$\|\mathbf{g}_k\| \leq (|\theta_k| + \|\mathbf{A}\|) \, \|\mathbf{q}_k\| \leq \left(1 + \frac{\|\mathbf{p}_k\| \, \|\mathbf{q}_k\|}{|\mathbf{p}_k^T \mathbf{q}_k|}\right) \|\mathbf{A}\| \, \|\mathbf{q}_k\|.$$

Substituting these formulas into (6.5), we arrive at the following result.

THEOREM 6.1. *The forward Euler iterates* (6.3)–(6.4) *for the two-sided dynamical system* (6.1) *satisfy*

$$\frac{|\mathbf{q}_{j+1}^T \mathbf{p}_{j+1} - \mathbf{q}_0^T \mathbf{p}_0|}{|\mathbf{q}_0^T \mathbf{p}_0|} \leq h^2 \frac{\|\mathbf{A}\|^2}{|\mathbf{q}_0^T \mathbf{p}_0|} \sum_{k=0}^{j} \left(1 + \frac{\|\mathbf{q}_k\| \, \|\mathbf{p}_k\|}{|\mathbf{q}_k^T \mathbf{p}_k|}\right)^2 \|\mathbf{q}_k\| \, \|\mathbf{p}_k\|. \tag{6.6}$$

This bound implies that the departure from the manifold is proportional to the square of the step size, and involves the secants of the angles formed by $\mathbf{q}_k$ and $\mathbf{p}_k$,

$k = 0, \ldots, j$, as well as the norms of $\mathbf{q}_k$ and $\mathbf{p}_k$. Moreover, unless the cosines of the angles between $\mathbf{q}_k$ and $\mathbf{p}_k$ are bounded away from zero, there does not exist a step size h such that all iterates remain near the quadratic manifold. The proof of the theorem demonstrates that the secant of the angle is at least as large as the normalized residuals. Numerical experiments indicate that these bounds are descriptive.

A similar result holds for the one-sided dynamical system with preconditioning (6.2), with forward Euler discretization given by

$$\mathbf{p}_{j+1} = \mathbf{p}_j + h\mathbf{N}^{-1}\mathbf{f}_j, \tag{6.7}$$

where now $\mathbf{f}_j = \mathbf{f}(\mathbf{p}_j, \mathbf{p}_j)$. Henceforth we assume that $\mathbf{N}$ is symmetric and invertible, which, as seen in the introduction, ensures that solutions of the continuous system reside on an invariant manifold $\mathbf{p}^T\mathbf{N}\mathbf{p} = \text{constant}$. At each time step, the discrete iteration incurs a local departure from that manifold of

$$e_{j+1} := \mathbf{p}_{j+1}^T\mathbf{N}\mathbf{p}_{j+1} - \mathbf{p}_j^T\mathbf{N}\mathbf{p}_j = h^2\mathbf{f}_j^T\mathbf{N}^{-1}\mathbf{f}_j.$$

Hence if $\mathbf{N}^{-1}$ is additionally positive definite (e.g., $\mathbf{N}^{-1} = \mathbf{I}$), the drift is monotone increasing—an important property for the forthcoming convergence theory.

When $\mathbf{N}$ is positive definite, we can define vector norms

$$\|\mathbf{z}\|_{\mathbf{N}^{-1}}^2 := \mathbf{z}^T\mathbf{N}^{-1}\mathbf{z}, \qquad \|\mathbf{z}\|_{\mathbf{N}}^2 := \mathbf{z}^T\mathbf{N}\mathbf{z}$$

(which in turn induce matrix norms), with $\|\mathbf{z}\|_{\mathbf{N}^{-1}} \le \|\mathbf{N}^{-1}\|\|\mathbf{z}\|_{\mathbf{N}}$. Thus we write

$$e_{j+1} = h^2\|\mathbf{f}_j\|_{\mathbf{N}^{-1}}^2 \le h^2\|\mathbf{N}^{-1}\|^2\|\mathbf{f}_j\|_{\mathbf{N}}^2 = h^2\|\mathbf{N}^{-1}\|^2\|\mathbf{r}_j\|_{\mathbf{N}}^2\|\mathbf{p}_j\|_{\mathbf{N}}^2,$$

where we use the normalized residual $\mathbf{r}_j := \mathbf{f}_j/\|\mathbf{p}_j\|_{\mathbf{N}} = (\theta_j - \mathbf{A})\mathbf{p}_j/\|\mathbf{p}_j\|_{\mathbf{N}}$. Now consider the aggregate, global drift from the manifold:

$$\begin{aligned} d_{j+1} &:= \mathbf{p}_{j+1}^T\mathbf{N}\mathbf{p}_{j+1} - \mathbf{p}_0^T\mathbf{N}\mathbf{p}_0 \\ &= \sum_{k=1}^{j+1} e_k \le h^2\|\mathbf{N}^{-1}\|^2 \sum_{k=0}^{j} \|\mathbf{r}_k\|_{\mathbf{N}}^2(d_k + \|\mathbf{p}_0\|_{\mathbf{N}}^2). \end{aligned}$$

In particular, d_{j+1} is determined by the step size, the residual norms, and the growth in the norm of the iterates. For further simplification, choose some $M > 0$ such that $\|\mathbf{r}_k\|_{\mathbf{N}}^2 \le M$ for all $k = 0, \ldots, j$. One coarse (but j-independent) possibility is

$$M := \inf_{s\in\mathbb{R}} 4\|\mathbf{A} - s\|_{\mathbf{N}}^2 \ge \inf_{s\in\mathbb{R}} \|(\mathbf{A} - s) - (\theta_k - s)\|_{\mathbf{N}}^2 \ge \|\mathbf{r}_k\|_{\mathbf{N}}^2, \tag{6.8}$$

which is invariant to shifts in $\mathbf{A}$. (In terms of the Euclidean norm, we thus have $M \le 4\kappa(\mathbf{N}) \inf_{s\in\mathbb{R}} \|\mathbf{A} - s\|^2$, where $\kappa(\mathbf{N}) = \|\mathbf{N}\|\|\mathbf{N}^{-1}\|$.) Hence

$$d_{j+1} \le h^2 M\|\mathbf{N}^{-1}\|^2 \sum_{k=0}^{j}(d_k + \|\mathbf{p}_0\|_{\mathbf{N}})^2 = h^2 M\|\mathbf{N}^{-1}\|^2\Big((j+1)\|\mathbf{p}_0\|_{\mathbf{N}}^2 + \sum_{k=1}^{j} d_k\Big)$$

(since $d_0 = 0$). Thus if we define the sequence $\{\widehat{d}_k\}$ by

$$\widehat{d}_{j+1} = h^2 M\|\mathbf{N}^{-1}\|^2\Big((j+1) + \sum_{k=1}^{j} \widehat{d}_k\Big), \tag{6.9}$$

then the departure from the manifold obeys $d_{j+1} \le \widehat{d}_{j+1}\|\mathbf{p}_0\|_{\mathbf{N}}^2$. Equation (6.9) is a binomial recurrence whose solution can be written explicitly:

$$\widehat{d}_{j+1} = \sum_{k=1}^{j+1} \binom{j+1}{k} (h^2 M \|\mathbf{N}^{-1}\|^2)^k = (1 + h^2 M \|\mathbf{N}^{-1}\|^2)^{j+1} - 1.$$

THEOREM 6.2. *Let $\mathbf{N} \in \mathbb{R}^{n\times n}$ be symmetric and positive definite, and define M by* (6.8)*. Then the forward Euler iterates* (6.7) *for the preconditioned one-sided dynamical system* (6.2) *satisfy*

$$0 \le \frac{\mathbf{p}_{j+1}^T \mathbf{N} \mathbf{p}_{j+1} - \mathbf{p}_0^T \mathbf{N} \mathbf{p}_0}{\mathbf{p}_0^T \mathbf{N} \mathbf{p}_0} \le (1 + h^2 M \|\mathbf{N}^{-1}\|^2)^{j+1} - 1, \tag{6.10}$$

the upper bound being asymptotic to $(j+1)h^2\|\mathbf{N}^{-1}\|^2 M$ as $h \to 0$.

Note that a small eigenvalue of $\mathbf{N}$ results in a small time-step h. The bound also provides an estimate of a critical time step

$$h\sqrt{j+1} \lessapprox \frac{1}{\|\mathbf{N}^{-1}\|\sqrt{M}}$$

for forward Euler, limiting the departure from the quadratic manifold. Highly non-normal problems for which $\|\mathbf{A} - s\| \gg \max_k |\lambda_k - s|$ also result in tiny time-steps.

Theorem 6.2 is a striking result—independent of starting vectors, the drift in the iterates for a one-sided iteration from the quadratic manifold is $O(h^2)$ and non-decreasing in j, under mild restrictions. This monotonic departure from the manifold is exploited in the discrete convergence analysis to follow. So, although, explicit Runge–Kutta methods (including forward Euler) do not preserve quadratic invariants (see [11, Chapter IV]), the forward Euler iterates for the one-sided systems remain nearby. The reader is referred to [11, Chapter IV] for further information and references, including the use of projection to remain on the quadratic manifold.

6.2. Discrete convergence theory. Just as the local drift from the manifold at each iteration contributes to the global drift, so local truncation errors committed by each step of an ODE solver aggregate into a global error. How does this accumulated error affect convergence of the discrete method as we compute $\mathbf{p}_j$ with $j \to \infty$? In this section, we seek conditions that will ensure that the discrete preconditioned one-sided iteration (6.2) converges to the same eigenvector as the continuous system.

Suppose $\mathbf{A} \in \mathbb{R}^{n\times n}$ has a simple eigenvalue λ_1 strictly to the left of all other eigenvalues (and hence real). Via a unitary transformation of coordinates, we write

$$\mathbf{A} = \begin{bmatrix} \lambda_1 & \mathbf{d}^T \\ \mathbf{0} & \mathbf{C} \end{bmatrix} \tag{6.11}$$

Let $\mathbf{x}_1$ and $\mathbf{y}_1$ denote unit-length right and left eigenvectors associated with λ_1; in these coordinates we can take $\mathbf{x}_1 = [1, 0, \ldots, 0]^T$. Theorems 4.3, 4.4, and 5.1 provide conditions under which the solution $\mathbf{p}(t)$ of the continuous system converges in angle to the eigenvector $\mathbf{x}_1$ (e.g., if $\mathbf{N} = \mathbf{I}$ and $\mathbf{y}_1^T \mathbf{p}_0 \ne 0$). Even when $\mathbf{p}_0$ satisfies such conditions, the discrete iteration can deflate the desired eigenvector, at which point convergence becomes impossible. One can write the iterate at step $k+1$ as

$$\mathbf{p}_{k+1} = \prod_{j=0}^{k} \varphi_j(\mathbf{A}) \mathbf{p}_0$$

for linear factors $\varphi_j(z) = 1 + h(\theta_j - z)$, with failure now equivalent to one of these polynomials having λ_1 as a root. Examples are simple to construct: for any *fixed* $h > 0$, set

$$\mathbf{A} = \begin{bmatrix} 0 & -1 - 2/h \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{p}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

so $\theta_0 = -1/h$, $\varphi_0(0) = 0$ and $\mathbf{p}_1 = [h+2, -h]^T$ is an eigenvector for $\lambda_2 = 1$. Note that $\varphi_j(\lambda_1) = 1+h(\theta_j-\lambda_1) = 0$ implies that $\theta_j-\lambda_1 = -1/h < 0$, and this is impossible if $\mathbf{A}$ is normal. As h is reduced, complete deflation requires an increasing departure from normality. (The more sophisticated restarted Arnoldi algorithm exhibits a similar phenomenon; see [7].)

Under what circumstances can we guarantee convergence? To answer this question, we first review the conventional global error analysis for the forward Euler method; for details, see, e.g., [8, §1.3]. The first step begins with the exact solution at time $t = 0$: $\mathbf{p}_0 = \mathbf{p}(0)$. Each subsequent step introduces a local truncation error, while also magnifying the global error aggregated at previous steps. Suppose we wish to integrate for $t \in [0, \tau]$ with $\tau = kh$ for some integer k. With the local truncation error at each step is bounded by

$$T_h := \max_{0 \le t \le \tau} \tfrac{1}{2} h \|\ddot{\mathbf{p}}(t)\|,$$

one can show that

$$\|\mathbf{p}_k - \mathbf{p}(\tau)\| \le \frac{T_h}{L}\big(e^{\tau L} - 1\big), \tag{6.12}$$

where L is a Lipschitz constant for our differential equation; in Appendix A we show that $L = 10\|\mathbf{N}^{-1}\|\|\mathbf{A}\|$ will suffice. This expression for the global error captures an essential feature: for fixed τ, the fact that $T_h = O(h)$ implies that we can always select $h > 0$ sufficiently small as to make the difference between the forward Euler iterate $\mathbf{p}_{\tau/h}$ and the exact solution $\mathbf{p}(\tau)$ arbitrarily small. However, if we increase k with $h > 0$ *fixed*, the bound indicates an *exponential* growth in the error. To show that $\mathbf{p}_k$ converges (in angle) to an eigenvector as $k \to \infty$, further work is required. In this effort, the preservation of the quadratic invariant characterized in Theorem 6.2 plays an essential role.

Preconditioning significantly complicates the convergence theory. For simplicity, our analysis imposes the stringent requirement that, in the coordinates in which $\mathbf{A}$ takes the form (6.11), we have

$$\mathbf{N}^{-1} = \begin{bmatrix} \eta & \mathbf{0} \\ \mathbf{0} & \mathbf{M} \end{bmatrix} \tag{6.13}$$

in addition to the requirement that $\mathbf{N}^{-1}$ be symmetric and positive definite. The trivial off-diagonal blocks prevent the preconditioner from using the growing component of $\mathbf{p}_k$ in $\mathbf{x}_1$ to enlarge the component in the unwanted eigenspace. In what follows, $\kappa(\mathbf{N}) = \|\mathbf{N}\|\|\mathbf{N}^{-1}\|$ denotes the condition number of the preconditioner.

THEOREM 6.3. *Given* (6.11), (6.13), *and assumptions on* λ_1, $\mathbf{x}_1$, *and* $\mathbf{N}$ *established in the previous paragraphs, suppose that* $\mathbf{p}_0$ *is chosen so that the continuous dynamical system converges in angle to an eigenvector associated with the distinct, simple leftmost eigenvalue* λ_1 *(e.g.,* $\mathbf{y}_1^T\mathbf{p}_0 \neq 0$ *suffices if* $\mathbf{N} = \mathbf{I}$*). Furthermore, suppose there exists* $h > 0$ *for which*

$$\gamma := \|\mathbf{I} + h\mathbf{M}(\lambda_1 - \mathbf{C})\| \in [0, 1/\sqrt{\kappa(\mathbf{N})}). \tag{6.14}$$

Then after preliminary iteration with a sufficiently small time-step h_0, the forward Euler method with time-step h will converge (in angle) to the desired eigenvector:

$$\sin(\angle(\mathbf{p}_k, \mathbf{x}_1)) = O(\gamma^k). \tag{6.15}$$

Asymptotically, the Ritz value converges to λ at the same rate:

$$|\theta_k - \lambda| = O(\gamma^k), \tag{6.16}$$

which in the case $\mathbf{d} = \mathbf{0}$ improves to $|\theta_k - \lambda| = O(\gamma^{2k})$.

Proof. Denote the kth iterate by

$$\mathbf{p}_k = \begin{bmatrix} \alpha_k \\ \mathbf{b}_k \end{bmatrix}.$$

To show that $\sin(\angle(\mathbf{p}_k, \mathbf{x}_1)) \to 0$ as $k \to \infty$, we will show that $\|\mathbf{b}_k\| \to 0$ while $|\alpha_k|$ is bounded away from zero. The convergence of the forward Euler method at a fixed time $\tau \geq 0$, with the assumption the continuous system converges for the given $\mathbf{p}_0$ (as described in Sections 4–5), ensures that we can run the forward Euler iteration with a sufficiently small time-step that, after $k \geq 0$ iterations, $\|\mathbf{b}_k\|$ is sufficiently small that

$$\frac{\|\mathbf{b}_k\|^2 \|\lambda_1 - \mathbf{C}\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\|\mathbf{b}_k\| \|\mathbf{d}\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}} \leq \frac{\varepsilon}{h\|\mathbf{M}\|} \tag{6.17}$$

for some $\varepsilon \in [0, 1/\sqrt{\kappa(\mathbf{N})} - \gamma)$; here $\gamma \in [0, 1/\sqrt{\kappa(\mathbf{N})})$ and $h > 0$ are as in the statement of the theorem. Theorem 6.2 and the fact that $\mathbf{N}$ is symmetric positive definite imply that

$$\|\mathbf{p}_k\|^2 \;\geq\; \frac{1}{\|\mathbf{N}\|} \mathbf{p}_k^T \mathbf{N} \mathbf{p}_k \;\geq\; \frac{1}{\|\mathbf{N}\|} \mathbf{p}_{k-1}^T \mathbf{N} \mathbf{p}_{k-1} \;\geq\; \frac{1}{\kappa(\mathbf{N})} \|\mathbf{p}_{k-1}\|^2, \tag{6.18}$$

so $|\alpha_k|$ must be bounded away from zero as $\|\mathbf{b}_k\| \to 0$. Since

$$\theta_k = \frac{\lambda_1 \alpha_k^2 + \alpha_k \mathbf{d}^T \mathbf{b}_k + \mathbf{b}_k^T \mathbf{C} \mathbf{b}_k}{\alpha_k^2 + \|\mathbf{b}_k\|^2},$$

we have

$$\begin{aligned} |\theta_k - \lambda_1| &= \frac{|\lambda_1 \alpha_k^2 + \alpha_k \mathbf{d}^T \mathbf{b}_k + \mathbf{b}_k^T \mathbf{C} \mathbf{b}_k - \lambda_1(\alpha_k^2 + \mathbf{b}_k^T \mathbf{b}_k)|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} \\ &\leq \frac{|\mathbf{b}_k^T (\mathbf{C} - \lambda_1) \mathbf{b}_k|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{|\alpha_k| \|\mathbf{b}_k\| \|\mathbf{d}\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} \\ &\leq \frac{\|\mathbf{b}_k\|^2 \, \|\mathbf{C} - \lambda_1\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\|\mathbf{b}_k\| \|\mathbf{d}\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}}, \end{aligned} \tag{6.19}$$

where the last inequality uses the fact that $|\alpha_k| \leq \sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}$. Now the condition (6.17) implies that the Ritz value θ_k is sufficiently close to the eigenvalue λ_1:

$$|\theta_k - \lambda_1| \leq \frac{\varepsilon}{h\|\mathbf{M}\|}. \tag{6.20}$$

The next step of the iteration, with time-step $h > 0$ specified in the statement of the theorem, produces

$$\begin{bmatrix}\alpha_{k+1}\\ \mathbf{b}_{k+1}\end{bmatrix} = \mathbf{p}_{k+1} = \mathbf{p}_k + h\mathbf{N}^{-1}(\theta_k - \mathbf{A})\mathbf{p}_k = \begin{bmatrix}\alpha_k + \eta h((\theta_k - \lambda_1)\alpha_k - \mathbf{d}^T\mathbf{b}_k)\\ (\mathbf{I} + h\mathbf{M}(\theta_k - \mathbf{C}))\mathbf{b}_k\end{bmatrix}.$$

Adding zero in a convenient way gives

$$\begin{aligned}\|\mathbf{b}_{k+1}\| &= \|(\mathbf{I} + h\mathbf{M}(\lambda_1 - \mathbf{C}))\mathbf{b}_k + h(\theta_k - \lambda_1)\mathbf{M}\mathbf{b}_k\| \\ &\le \|\mathbf{I} + h\mathbf{M}(\lambda_1 - \mathbf{C})\|\|\mathbf{b}_k\| + h|\lambda_1 - \theta_k|\|\mathbf{M}\|\|\mathbf{b}_k\| \\ &\le (\gamma + \varepsilon)\|\mathbf{b}_k\|. \end{aligned} \tag{6.21}$$

In particular, this guarantees a fixed reduction in the component of the forward Euler iterate in the unwanted eigenspace. (The second inequality follows from condition (6.14) and bound (6.20).) After checking a few details, we shall see that this condition is the key to convergence.

We now show that the new Ritz value, θ_{k+1}, automatically satisfies the requirement (6.20) with the same $\varepsilon > 0$ and time-step. Repeating the calculation that culminated in (6.19), we obtain

$$|\theta_{k+1} - \lambda_1| \le \frac{\|\mathbf{b}_{k+1}\|^2 \|\mathbf{C} - \lambda_1\|}{\alpha_{k+1}^2 + \|\mathbf{b}_{k+1}\|^2} + \frac{\|\mathbf{d}\|\|\mathbf{b}_{k+1}\|}{\sqrt{\alpha_{k+1}^2 + \|\mathbf{b}_{k+1}\|^2}}.$$

Now we use (6.18), a consequence of the monotonic drift from the invariant manifold, to deduce that

$$\begin{aligned}|\theta_{k+1} - \lambda_1| &\le \frac{\kappa(\mathbf{N})(\gamma + \varepsilon)^2\|\mathbf{b}_k\|^2 \|\mathbf{C} - \lambda_1\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\sqrt{\kappa(\mathbf{N})}(\gamma + \varepsilon)\|\mathbf{d}\|\|\mathbf{b}_k\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}} \\ &\le \frac{\|\mathbf{b}_k\|^2 \|\mathbf{C} - \lambda_1\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\|\mathbf{d}\|\|\mathbf{b}_k\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}}, \end{aligned}$$

since $\gamma + \varepsilon < 1/\sqrt{\kappa(\mathbf{N})}$. The condition (6.17) then implies that

$$|\theta_{k+1} - \lambda_1| \le \frac{\varepsilon}{h\|\mathbf{M}\|},$$

which guarantees that the Ritz value cannot wander too far from λ_1. Furthermore, this bound allows us to repeat the argument resulting in (6.21) at future steps, giving

$$\|\mathbf{b}_{k+m}\| \le (\gamma + \varepsilon)^m\|\mathbf{b}_k\|$$

along with, via a slight modification of (6.18),

$$\begin{aligned}|\theta_{k+m} - \lambda_1| &\le \frac{\kappa(\mathbf{N})(\gamma + \varepsilon)^{2m}\|\mathbf{b}_k\|^2 \|\mathbf{C} - \lambda_1\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\sqrt{\kappa(\mathbf{N})}(\gamma + \varepsilon)^m\|\mathbf{d}\|\|\mathbf{b}_k\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}} \\ &\le \frac{\|\mathbf{b}_k\|^2 \|\mathbf{C} - \lambda_1\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\|\mathbf{d}\|\|\mathbf{b}_k\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}}. \end{aligned} \tag{6.22}$$

Thus $|\theta_{k+m} - \lambda_1| \le \varepsilon/(h\|\mathbf{M}\|)$ for all $m \ge 1$. As $\|\mathbf{b}_{k+m}\| \to 0$, the component in the desired eigenvector is bounded away from zero, as again a generalization of (6.18) gives

$$\|\mathbf{p}_{k+m}\| \ge \frac{1}{\sqrt{\kappa(\mathbf{N})}}\|\mathbf{p}_0\|.$$

Thus with $\mathbf{x}_1 = \mathbf{e}_1$, we have

$$\begin{aligned}\sin\angle(\mathbf{p}_{k+m}, \mathbf{x}_1) &= \min_{\xi} \frac{\|\xi\mathbf{p}_{k+m} - \mathbf{x}_1\|}{\|\mathbf{x}_1\|} = \min_{\xi} \left\| \begin{bmatrix} \xi\alpha_{k+m} - 1 \\ \xi\mathbf{b}_{k+m} \end{bmatrix} \right\| \\ &\le \frac{\|\mathbf{b}_{k+m}\|}{|\alpha_{k+m}|} \le (\gamma+\varepsilon)^m \frac{\|\mathbf{b}_k\|}{|\alpha_{k+m}|}\end{aligned}$$

where we have taken $\xi = \alpha_{k+m}^{-1}$ for the first inequality. As $|\alpha_{k+m}|$ is bounded away from zero, we have $\sin\angle(\mathbf{p}_{k+m}, \mathbf{x}_1) = O((\gamma+\varepsilon)^m)$ as $m \to \infty$. Since $\|\mathbf{b}_{k+m}\| \to 0$ as $m \to \infty$, we can take the ε used in (6.19) to be arbitrarily small as the iterations progress, giving the asymptotic rate given in (6.15). Similarly, from (6.22) we observe that the Ritz value converges as in (6.16). The $O(\gamma^m)$ term in that bound falls out if $\mathbf{d} = \mathbf{0}$. □

We now make several remarks concerning the previous theorem and its proof. (1) Given the special form of $\mathbf{A}$ and $\mathbf{N}$, the constant γ defined in (6.14) can be written in the more general form

$$\gamma = \|\mathbf{\Pi}_1(\mathbf{I} + h\mathbf{N}(\lambda_1 - \mathbf{A})\|,$$

where $\mathbf{\Pi}_1 = \mathbf{I} - \mathbf{x}_1\mathbf{x}_1^T$ is the orthogonal projector onto the undesired invariant subspace. (2) The condition (6.14) implies that as $\mathbf{N}$ gets increasingly ill-conditioned, our convergence theory requires it to become ever more effective. (3) A curiosity of condition (6.17) is that the requirement is more strict when convergence is slower, i.e., when γ is near $\kappa(\mathbf{N})^{-1/2}$. (4) One does not in general know whether θ_k falls to the left or right of λ_1. If $\mathbf{A}$ is normal, then as θ_k must fall the convex hull of its spectrum, and so $\theta_k \ge \lambda_1$; for nonnormal $\mathbf{A}$, it is possible that $\theta_k < \lambda_1$. (5) The proof of the theorem exploits the monotonic drift from the manifold described by Theorem 6.2. This drift is easily monitored, so providing a useful (and cheap) check on convergence of the iteration during computation. If this drift reaches a point where it is not small, projection to the quadratic manifold is easily undertaken; see [11, Chapter IV] for further information.

Theorem 6.3 considers the general case of nonsymmetric $\mathbf{A}$ and a somewhat stringent notion of preconditioning. For the important special case of symmetric positive definite $\mathbf{A}$, Knyazev and Neymeyr [16] provide convergence estimates (and review much literature) for the one-sided forward Euler discretization (6.3). They provide rates of convergence given a symmetric positive definite preconditioner $\mathbf{N}$ for $\mathbf{A}$. However, a connection with dynamical systems is not made and instead optimization is applied to the Rayleigh quotient.

If $\mathbf{M} = \mathbf{I}$, and $\mathbf{C}$ is normal (which is possible even if $\mathbf{A}$ itself is not normal due to $\mathbf{d} \ne \mathbf{0}$) with spectrum given by $\sigma(\mathbf{C}) = \{\lambda_2, \ldots, \lambda_n\}$, we have

$$\gamma := \max_{i=2,\ldots,n} |1 + h(\lambda_1 - \lambda_i)|,$$

and we can apply the equioscillation theorem to determine the optimal $h > 0$. In particular, if all the eigenvalues are real (thus $\mathbf{C}$ is symmetric) and $\lambda_2 \le \lambda_3 \le \cdots \le \lambda_n$,

then the best h must give

$$1 + h(\lambda_1 - \lambda_2) = -1 - h(\lambda_1 - \lambda_n).$$

This can be solved to obtain $h = 2/(\lambda_2 + \lambda_n - 2\lambda_1)$, from which we compute

$$\gamma = \frac{\lambda_n - \lambda_2}{\lambda_n + \lambda_2 - 2\lambda_1}.$$

Notice that this agrees with the convergence rate of the power method applied to $\mathbf{A} - \sigma\mathbf{I}$ for the optimal shift $\sigma = \frac{1}{2}(\lambda_2 + \lambda_n)$ to the leftmost eigenvector $\mathbf{x}_1$; see, e.g., [32, p. 572]. With the optimal choice of h, the forward Euler method recovers the convergence rate of an optimally shifted power method to $\mathbf{x}_1$.

Again, suppose that $\mathbf{M} = \mathbf{I}$, so that $\gamma = \gamma(h) \to 1$ as $h \to 0$. However, this limit need not be approached from below; that is, for some matrices $\mathbf{C}$ we will have $\gamma(h) > 1$ for all h sufficiently small. The behavior of γ in this limit bears a close connection to the *logarithmic norm* of $\lambda_1 - \mathbf{C}$, which is defined as

$$\beta(\lambda_1 - \mathbf{C}) := \lim_{h \downarrow 0} \frac{\|\mathbf{I} + h(\lambda_1 - \mathbf{C})\| - 1}{h};$$

see, e.g., [20],[29, Chap. 17]. In particular, $\gamma(h) < 1$ for all sufficiently small $h > 0$ provided $\beta(\lambda_1 - \mathbf{C}) < 0$. One can show that the logarithmic norm of a matrix coincides with the numerical abscissa, that is, the real part of the rightmost point in the numerical range:

$$\begin{aligned}\beta(\lambda_1 - \mathbf{C}) &= \max_{\mathbf{v} \in \mathbb{C}^{n-1}, \|\mathbf{v}\| = 1} \text{Re } \mathbf{v}^*(\lambda_1 - \mathbf{C})\mathbf{v} \\ &= \max\{\eta : \eta \in \sigma(\tfrac{1}{2}((\lambda_1 - \mathbf{C}) + (\lambda_1 - \mathbf{C}^T))\},\end{aligned}$$

see, e.g., [29, Theorem 17.4]. When is $\gamma(h) > 1$? That is, for what matrices can we not apply our convergence theory by taking h arbitrarily small? It is equivalent to find requirements on $\mathbf{C}$ that ensure $\beta(\lambda_1 - \mathbf{C}) < 0$. From the above analysis we see that

$$\beta(\lambda_1 - \mathbf{C}) = \lambda_1 - \min_{\mathbf{v} \in \mathbb{C}^{n-1}, \|\mathbf{v} = 1\|} \text{Re } \mathbf{v}^*\mathbf{C}\mathbf{v},$$

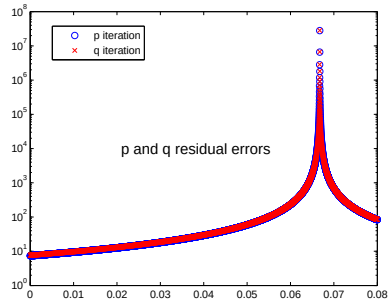
from which we conclude the following.

LEMMA 6.4. *The logarithmic norm $\beta(\lambda_1 - \mathbf{C}) < 0$ (equivalently, $\gamma < 1$ for all sufficiently small $h > 0$) if and only if the numerical range of $\mathbf{C}$ does not include λ_1.*

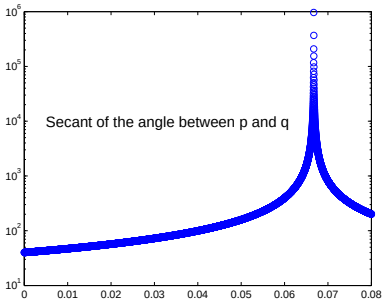
6.3. Numerical experiments. In this section we investigate the Theorems 6.1 and 6.3 through several computational examples.

Figure 6.1 samples the flow in h-intervals given by Theorem 4.1 for the tridiagonal matrix

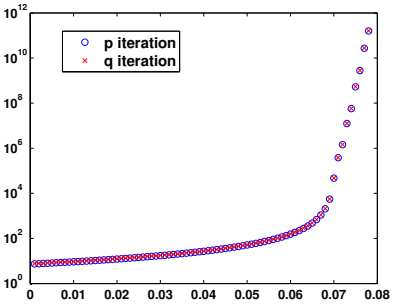
$$\mathbf{T}_\rho^n \equiv \begin{bmatrix} 2 & -1+\rho & & 0 \\ -1-\rho & 2 & \ddots & \\ & \ddots & \ddots & -1+\rho \\ 0 & & -1-\rho & 2 \end{bmatrix} \in \mathbb{R}^{n \times n}$$



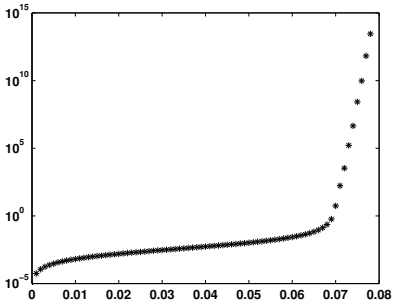
(a) $\|\mathbf{f}_j\|$ and $\|\mathbf{g}_j\|, h = 10^{-5}$



(b) $(\|\mathbf{p}\|\|\mathbf{q}\|)/|\mathbf{q}^T\mathbf{p}|, h = 10^{-5}$



(c) $\|\mathbf{f}_j\|$ and $\|\mathbf{g}_j\|, h = 10^{-3}$



(d) $\left|\frac{\mathbf{q}_j^T\mathbf{p}_j}{\mathbf{q}_0^T\mathbf{p}_0} - 1\right|, h = 10^{-3}$

FIG. 6.1. *Sampled flow for* $\mathbf{T}_\rho^{100}$ *where* $\rho = 1/(20 \cdot 101)$ *over two time intervals. The horizontal axis measures time.*

where $n = 100$ and $\rho = 1/(20(n+1))$. The eigenvalues are all real and the condition number of the matrix of eigenvectors is modest. The experiments of Figure 6.1 use the same starting vectors.

Figure 6.1(a),(b) show that at $t \approx .067$, a cusp develops indicating that a pole as given by $\pi(t)$ of Theorem 4.1 is encountered by the discrete flow. Figure 6.1(c),(d) displays the discrete flow associated with a forward Euler time integrator with a time step of $h = 10^{-3}$. As expected, when the iterates depart from the quadratic manifold, the residuals explode in size. One can also show that the secant of the angle between $\mathbf{p}_j$ and $\mathbf{q}_j$, and the norms of $\mathbf{p}_j$ and $\mathbf{q}_j$ also explode, demonstrating that Theorem 6.1 is descriptive.

Decreasing the time-step h does not avoid the error—in fact, the time at which the explosive growth occurs is independent of the time-step because of the onset of incurable breakdown associated with the continuous dynamical system. In contrast to the latter, the discrete dynamical system cannot simply step over the pole associated with continuous dynamical system. The special case of Theorem 4.2 aside, these results are typical and do not depend on specially engineered starting vectors. We also implemented the symplectic Euler method (that preserves quadratic invariants) and forward Euler combined with a projection. The results obtained are consistent

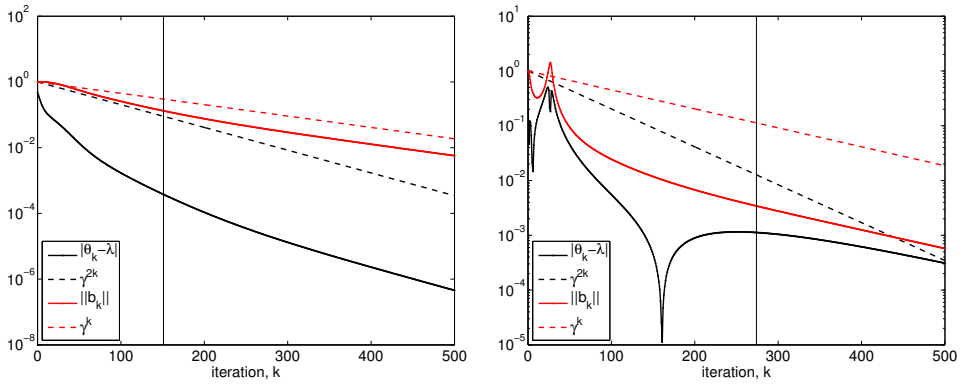


FIG. 6.2. *Computational confirmation of Theorem 6.3 for a normal matrix (left) and a non-normal matrix (right), both with* $\mathbf{N} = \mathbf{I}$.

with forward Euler (as displayed in Figure 6.1)(c),(d). In contrast, the one-sided discretized forward Euler iterations converge to the left eigenvalue and associated eigenvector.

Next we investigate the convergence analysis described in Theorem 6.3 for a simple example with $\mathbf{N} = \mathbf{I}$. Let $\mathbf{A}$ be the matrix with $a_{j,j} = (j-1)/(N-1)$ for $j = 1, \ldots, N$, and all other entries equal to zero except perhaps for the vector $\mathbf{d}^T$ in entries 2 through N of the first row; cf. (6.11). The plots in Figure 6.2 use $N = 64$, comparing $\mathbf{d}^T = \mathbf{0}$ (left) and $\mathbf{d}^T = [1, \ldots, 1]$ (right). In both cases we take $h = 1/2$, for which (6.14) gives $\gamma = 0.992\ldots \in [0, 1)$ as required. We take $\mathbf{p}_0$ to be the same randomly-generated unit vector in both cases. This initial vector does not satisfy (6.17), but this condition is eventually met after a number of iterations, denoted by the vertical line in each plot. For the normal case in the left plot, $\|\mathbf{b}_k\|$ converges like γ^k, while the error in the Ritz value $|\theta_k - \lambda_1|$ converges like γ^{2k} as predicted. The nonnormality induced by the $\mathbf{d}$ vector spoils this convergence for the Ritz value, as seen in the right plot; now both $\|\mathbf{b}_k\|$ and $|\theta_k - \lambda_1|$ converge like γ^k, consistent with Theorem 6.3. The spikes in the latter plot correspond to points where the Ritz value θ_k crossed over the desired eigenvalue λ_1, something only possible for nonnormal iterations.

7. Summary. This paper demonstrates the fruitful relationship between several nonlinear dynamical systems and certain simple preconditioned eigensolvers for non-symmetric eigenvalue problems. Properties of the continuous-time systems, such as system invariants and the asymptotic behavior of the exact solution, can inform the convergence theory for practical algorithms derived from discretizations, as we illustrate with Theorem 6.1 for the forward Euler discretization. Generalizations to more sophisticated discretizations, along with relaxation of the stringent requirements on the preconditioner in Theorem 6.1, are natural avenues for future research.

Acknowledgements. We thank Pierre-Antoine Absil, Kyle Gallivan, Anthony Kellems, Christian Lubich, and Qiang Ye for their helpful comments.

Appendix A. Lipschitz constant for Euler's method. To apply the standard convergence theory for the forward Euler method applied to the system

$$\dot{\mathbf{p}} = \mathbf{N}^{-1}(\theta(\mathbf{p})\mathbf{p} - \mathbf{A}\mathbf{p}),$$

we seek a constant $L > 0$ such that

$$\|\mathbf{N}^{-1}(\theta(\mathbf{u})\mathbf{u} - \mathbf{A}\mathbf{u}) - \mathbf{N}^{-1}(\theta(\mathbf{v})\mathbf{v} - \mathbf{A}\mathbf{v})\| \le L\,\|\mathbf{u} - \mathbf{v}\|$$

for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. First we note that

$$\|(\theta(\mathbf{u})\mathbf{u} - \mathbf{A}\mathbf{u}) - (\theta(\mathbf{v})\mathbf{v} - \mathbf{A}\mathbf{v})\| \le \|\theta(\mathbf{u})\mathbf{u} - \theta(\mathbf{v})\mathbf{v}\| + \|\mathbf{A}\|\|\mathbf{u} - \mathbf{v}\|.$$

We focus attention on the first term on the right:

$$\begin{aligned}\|\theta(\mathbf{u})\mathbf{u} - \theta(\mathbf{v})\mathbf{v}\| &\le \|\theta(\mathbf{u})\mathbf{u} - \theta(\mathbf{v})\mathbf{u} + \theta(\mathbf{v})\mathbf{u} - \theta(\mathbf{v})\mathbf{v}\| \\ &\le |\theta(\mathbf{u}) - \theta(\mathbf{v})|\|\mathbf{u}\| + |\theta(\mathbf{v})|\|\mathbf{u} - \mathbf{v}\| \\ &\le |\theta(\mathbf{u}) - \theta(\mathbf{v})|\|\mathbf{u}\| + \|\mathbf{A}\|\|\mathbf{u} - \mathbf{v}\|. \qquad \text{(A.1)}\end{aligned}$$

(In this last inequality and others that follow, we neglect the opportunity to take tighter bounds that would lead to smaller constants but greater analytical complexity.)

We next we need to bound $|\theta(\mathbf{u}) - \theta(\mathbf{v})|\|\mathbf{u}\|$ in terms of $\|\mathbf{u} - \mathbf{v}\|$. For convenience (assuming neither $\mathbf{u}$ nor $\mathbf{v}$ is zero), define the unit vectors $\widehat{\mathbf{u}} = \mathbf{u}/\|\mathbf{u}\|$ and $\widehat{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|$, with $\boldsymbol{\varepsilon} = \widehat{\mathbf{v}} - \widehat{\mathbf{u}}$, so that

$$\begin{aligned}|\theta(\mathbf{u}) - \theta(\mathbf{v})| &= |\widehat{\mathbf{u}}^T\mathbf{A}\widehat{\mathbf{u}} - \widehat{\mathbf{v}}^T\mathbf{A}\widehat{\mathbf{v}}| \\ &= |\widehat{\mathbf{u}}^T\mathbf{A}\widehat{\mathbf{u}} - \widehat{\mathbf{u}}^T\mathbf{A}\widehat{\mathbf{u}} - \boldsymbol{\varepsilon}^T\mathbf{A}\widehat{\mathbf{u}} - \widehat{\mathbf{u}}^T\mathbf{A}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T\mathbf{A}\boldsymbol{\varepsilon}| \\ &\le 2\|\boldsymbol{\varepsilon}\|\|\mathbf{A}\| + \|\boldsymbol{\varepsilon}\|^2\|\mathbf{A}\|. \qquad \text{(A.2)}\end{aligned}$$

Now note that

$$\|\boldsymbol{\varepsilon}\| = \|\widehat{\mathbf{v}} - \widehat{\mathbf{u}}\| = \frac{\Big\|\|\mathbf{u}\|\mathbf{v} - \|\mathbf{v}\|\mathbf{v} + \|\mathbf{v}\|\mathbf{v} - \|\mathbf{v}\|\mathbf{u}\Big\|}{\|\mathbf{u}\|\|\mathbf{v}\|} \le \frac{\Big|\|\mathbf{u}\| - \|\mathbf{v}\|\Big|}{\|\mathbf{u}\|} + \frac{\|\mathbf{u} - \mathbf{v}\|}{\|\mathbf{u}\|}.$$

Apply the triangle inequality to obtain $\Big|\|\mathbf{u}\| - \|\mathbf{v}\|\Big| \le \|\mathbf{u} - \mathbf{v}\|$, from which we conclude

$$\|\boldsymbol{\varepsilon}\| \le \frac{2}{\|\mathbf{u}\|}\|\mathbf{u} - \mathbf{v}\|. \qquad \text{(A.3)}$$

Since $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{v}}$ are unit vectors, we alternatively have the coarse bound $\|\boldsymbol{\varepsilon}\| = \|\widehat{\mathbf{u}} - \widehat{\mathbf{v}}\| \le 2$, which we can apply to (A.2) to obtain

$$\begin{aligned}|\theta(\mathbf{u}) - \theta(\mathbf{v})| &\le 2\|\boldsymbol{\varepsilon}\|\|\mathbf{A}\| + \|\boldsymbol{\varepsilon}\|^2\|\mathbf{A}\| \\ &\le 2\|\boldsymbol{\varepsilon}\|\|\mathbf{A}\| + 2\|\boldsymbol{\varepsilon}\|\|\mathbf{A}\| \;=\; 4\|\mathbf{A}\|\|\boldsymbol{\varepsilon}\|.\end{aligned}$$

Now using (A.3), the bound first bound on $\|\boldsymbol{\varepsilon}\|$,

$$|\theta(\mathbf{u}) - \theta(\mathbf{v})| \le 8\frac{\|\mathbf{A}\|}{\|\mathbf{u}\|}\|\mathbf{u} - \mathbf{v}\|.$$

Substituting this bound into (A.1) gives

$$\|\theta(\mathbf{u})\mathbf{u} - \theta(\mathbf{v})\mathbf{v}\| \le 9\|\mathbf{A}\|\|\mathbf{u} - \mathbf{v}\|,$$

and finally we arrive at the Lipschitz constant

$$\|\mathbf{N}^{-1}(\theta(\mathbf{u})\mathbf{u}-\mathbf{A}\mathbf{u})-\mathbf{N}^{-1}(\theta(\mathbf{v})\mathbf{v}-\mathbf{A}\mathbf{v})\| \le 10\|\mathbf{N}^{-1}\|\|\mathbf{A}\|\|\mathbf{u}-\mathbf{v}\|.$$

Thus we define

$$L = 10\|\mathbf{N}^{-1}\|\|\mathbf{A}\|. \tag{A.4}$$

The Rayleigh quotient $\theta(\mathbf{p})$ is undefined in the case that $\mathbf{p}=\mathbf{0}$. However, as $\|\mathbf{p}\| \to 0$, we have that $\|\theta(\mathbf{p})\mathbf{p}-\mathbf{A}\mathbf{p}\| \to 0$, and this motivates the definition that $\theta(\mathbf{p})\mathbf{p}-\mathbf{A}\mathbf{p}=\mathbf{0}$ if $\mathbf{p}=\mathbf{0}$.

The above analysis excludes the case that $\mathbf{u}=\mathbf{0}$ and/or $\mathbf{v}=\mathbf{0}$, but with our definition of this singular case we have, e.g., if $\mathbf{u}=\mathbf{0}$, that

$$\|(\theta(\mathbf{u})\mathbf{u}-\mathbf{A}\mathbf{u})-(\theta(\mathbf{v})\mathbf{v}-\mathbf{A}\mathbf{v})\| = \|\theta(\mathbf{v})\mathbf{v}-\mathbf{A}\mathbf{v})\| \le 2\|\mathbf{A}\|\|\mathbf{v}\| \le 10\|\mathbf{A}\|\|\mathbf{u}-\mathbf{v}\|,$$

and obviously if $\mathbf{u}=\mathbf{v}=\mathbf{0}$, we have

$$\|\theta(\mathbf{u})\mathbf{u}-\mathbf{A}\mathbf{u})-(\theta(\mathbf{v})\mathbf{v}-\mathbf{A}\mathbf{v})\| = 0 = 10\|\mathbf{A}\|\|\mathbf{u}-\mathbf{v}\|.$$

Hence, the Lipschitz constant (A.4) holds for all $\mathbf{u}$ and $\mathbf{v}$.

REFERENCES

[1] P.-A. Absil, *Continuous-time systems that solve computational problems*, Int. J. Uncov. Comput., 2 (2006), pp. 291–304.

[2] V. I. Arnold, *Ordinary Differential Equations*, Springer-Verlag, Berlin, 3rd ed., 1992.

[3] W. Bao and Q. Du, *Computing the ground state solution of Bose–Einstein condensates by a normalized gradient flow*, SIAM J. Sci. Comp., 25 (2004), pp. 1674–1697.

[4] R. Car and M. Parrinello, *Unified approach for molecular dynamics and density functional theory*, Phys. Rev. Lett., 55 (1985), pp. 2471–2474.

[5] J. Carr, *Applications of Centre Manifold Theory*, Springer-Verlag, New York, 1981.

[6] M. T. Chu, *On the continuous realization of iterative processes*, SIAM Review, (1988), pp. 375–387.

[7] M. Embree, *The Arnoldi eigenvalue iteration with exact shifts can fail*, SIAM J. Matrix Anal. Appl. To appear.

[8] C. W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[9] G. H. Golub and L.-Z. Liao, *Continuous methods for extreme and interior eigenvalue problems*, Linear Algebra Appl., 415 (2006), pp. 31–51.

[10] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[11] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer-Verlag, Berlin, 2nd ed., 2006.

[12] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*, Springer, London, 1994.

[13] W. Kahan, B. N. Parlett, and E. Jiang, *Residual bounds on approximate eigensystems of nonnormal matrices*, SIAM J. Num. Anal., 19 (1982), pp. 470–484.

[14] A. V. Knyazev, *Preconditioned eigensolvers—an oxymoron?*, Elec. Trans. Numer. Anal., 7 (1998), pp. 104–123.

[15] A. V. Knyazev and K. Neymeyr, *Efficient solution of symmetric eigenvalue problems using multigrid preconditioners in the locally optimal block conjugate gradient method*, Elec. Trans. Numer. Anal., 7 (2003), pp. 38–55.

[16] A. V. Knyazev and K. Neymeyr, *A geometric theory for preconditioned inverse iteration. III: A short and sharp convergence estimate for generalized eigenvalue problems*, Linear Algebra Appl., 358 (2003), pp. 95–114.

[17] R. B. Lehoucq and A. J. Salinger, *Large-scale eigenvalue calculations for stability analysis of steady flows on massively parallel computers*, Int. J. Num. Methods in Fluids, 36 (2001), pp. 309–327.

[18] R. Mahony and P.-A. Absil, *The continuous time Rayleigh quotient flow on the sphere*, Linear Algebra Appl., 368 (2003), pp. 343–357.
[19] T. Nanda, *Differential equations and the QR algorithm*, SIAM J. Num. Anal., 22 (1985), pp. 310–321.
[20] O. Nevanlinna, *Convergence of Iterations for Linear Equations*, Birkhäuser, Basel, 1993.
[21] E. E. Osborne, *On pre-conditioning of matrices*, J. ACM, 7 (1960), pp. 338–345.
[22] B. N. Parlett, *The Symmetric Eigenvalue Problem*, no. 20 in Classics in Applied Mathematics, SIAM, 1998. Amended reprint of 1980 Prentice-Hall edition.
[23] B. N. Parlett and C. Reinsch, *Balancing a matrix for calculation of eigenvalues and eigenvectors*, Num. Math., 13 (1969), pp. 293–304.
[24] M. C. Payne, M. P. Teeter, D. C. Allan, T. Arias, and J. Joannopoulos, *Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients*, Rev. Mod. Phys, 64 (1992), pp. 1045–1097.
[25] B. T. Polyak, *Introduction to Optimization*, Translation Series in Mathematics and Engineering, Optimization Software, Inc., New York, 1987.
[26] Y. Saad, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl., 34 (1980), pp. 269–295.
[27] G. W. Stewart, *Matrix Algorithms*, vol. II: Eigensystems, SIAM, Philadelphia, 2001.
[28] W. W. Symes, *The QR algorithm and scattering for the finite nonperiodic Toda lattice*, Physica D, 4 (1982), pp. 275–280.
[29] L. N. Trefethen and M. Embree, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, Princeton, 2005.
[30] J. S. Warsa, T. A. Wareing, J. E. Morel, J. M. McGhee, and R. B. Lehoucq, *Krylov subspace iterations for deterministic k-eigenvalue calculations*, Nuc. Sci. Engrg., 147 (2004), pp. 26–42.
[31] D. S. Watkins, *Isospectral flows*, SIAM Review, 26 (1984), pp. 379–391.
[32] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.